# Zur Messung im Wirtschafts- und Sozialwissenschaftlichen Umfeld

## Eine empirische Studie zur Anzahl der Antwortkategorien bei einer Ratingskala

Dissertation

zur Erlangung des Doktorgrades an der

Wirtschaftswissenschaftlichen Fakultät

der Universität Augsburg

vorgelegt von

Alexander Spatz

2008

| | |
|---|---|
| Erstgutachter: | Prof. Dr. Otto Opitz |
| Zweitgutachter: | Prof. Dr. Robert Klein |
| Vorsitzender der Disputation: | Prof. Dr. Axel Tuma |

Tag der mündlichen Prüfung: 6. Mai 2008

# Inhaltsverzeichnis

## Kapitel 1

## Einleitung



## Kapitel 2

## Grundlagen der Messtheorie



## Kapitel 3

## Erhebung und Messung im wirtschaftswissenschaftlichen Kontext

## Kapitel 4

## Beantwortungseffekte und ihre Ursachen

# Abbildungsverzeichnis

# Tabellenverzeichnis

# Kapitel 1

# Einleitung

## 1.1 Problemstellung und Zielsetzung

Durch den hohen Konkurrenzdruck im internationalen Wettbewerb und einer stetig anhaltenden Marktsättigung wird es für Unternehmen immer wichtiger, die Bedürfnisse spezieller Kundengruppen und anderer Beteiligter in die Markt- und Absatzpolitik mit einzubeziehen. Nur so können Ziele und Strategien definiert werden, die die Unternehmensexistenz langfristig sichern. Getreu dem Motto des amerikanischen Sozialforschers ERNEST DICHTER (1907-1991) „Wenn ich Hundefutter verkaufen will, muss ich erst einmal die Rolle des Hundes übernehmen; denn nur der Hund allein weiß ganz genau, was Hunde wollen“, scheint es von entscheidender Bedeutung für die Wirtschafts- und Sozialwissenschaften zu sein, Erkenntnisse über das Denken, Fühlen und die Einstellung beispielsweise von Konsumenten zu Produkten, Marken oder Image zu gewinnen.

Insbesondere durch die Entwicklung der elektronischen Datenverarbeitung haben sich umfangreiche empirische Erhebungen als das übliche Instrument zur Gewinnung von Informationen über einzelne Personen oder Personenkreise entwickelt. Zumeist erhalten Unternehmen ihre Kenntnisse über die direkte Befragung Ihrer Konsumenten auf Grundlage von Fragebögen. Dabei werden häufig Skalierungsverfahren, wie die Ratingskala, eingesetzt, mit deren Hilfe z.B. einem Einstellungsobjekt bestimmte Messwerte zugeordnet werden können.

Die Informationen müssen möglichst fehlerfrei und lückenlos erfasst und berechnet werden, denn die letztendlichen Entscheidungen können immer nur so gut sein, wie die Informationen, auf denen sie basieren. Die Auswertung der mit Hilfe der Ratingskalen erhobenen Daten wurde in der Vergangenheit ausreichend erforscht und ist daher weitgehend fehlerfrei umsetzbar.

Doch selbst „[…] anspruchsvolle und komplexe mathematische Methoden können unsinnige Zuordnungen von Zahlen zu Ereignissen nicht mehr verbessern“ (KREPPNER, 1975, S. 13). Probleme entstehen also nicht zwangsläufig nur durch die Weiterverarbeitung der erfassten Daten, sondern auch durch Verzerrungen, die bereits vorher, bei der Daten*erfassung* an sich, auftreten können. Die Ausgestaltung der Fragebögen – und damit letztendlich der Ratingskalen – muss folglich sinnvoll vorgenommen werden, da darauf die wirksame Entwicklung von unternehmerischen Strategien und Maßnahmen beruht.

Die zentrale Frage lautet daher: Wie kann die Fehlinterpretation von Probandeninformationen am besten durch Unterstützung geeigneter Ratingskalen vermieden werden?

Ziel dieser Arbeit ist es, unter Berücksichtigung der bisherigen Forschungsergebnisse und auf der Grundlage einer selbst durchgeführten empirischen Studie herauszufinden, wie viele Antwortkategorien bei der Verwendung von Ratingskalen zweckmäßig sind, um Verzerrungen bei der Datenerfassung zu minimieren. Des Weiteren sollen Auswirkungen psychologischer Prozesse auf die Befragungsergebnisse ermittelt werden, um so deren Einfluss zu verringern.

Ausgangspunkt der Überlegungen ist die Frage, *ob* etwas überhaupt messbar ist. Darauf aufbauend gilt es zu zeigen, *wie* etwas messbar ist. Des Weiteren spielen die *psychologischen Prozesse,* die Probanden bei der Beantwortung von Fragen durchlaufen, eine wichtige Rolle für die Qualität der Messergebnisse. Da aktuelle Studien zur Ausgestaltung von Ratingskalen sehr unterschiedliche Meinungen zu dem Thema „Anzahl der Antwortkategorien“ aufweisen, soll eine eigene Studie Aufschluss darüber geben, wie viele Antwortkategorien bei gewissen Fragestellungen sinnvoll sind.

## 1.2 Vorgehen und methodischer Aufbau

Nach der allgemeinen Einleitung befasst sich **Kapitel 2** mit der *Messtheorie* als konzeptionelle Grundlage für Messungen im wirtschafts- und sozialwissenschaftlichen Bereich. Wesentliche Voraussetzung für eine nähere Betrachtung ist in Abschnitt 2.1 die Definition einiger grundlegender *Begriffe der Messtheorie*, insbesondere auch des Begriffs der Messung selbst. Punkt 2.2 beschäftigt sich mit dem *Repräsentationsproblem*. Bei einer Messung werden den Auskunftspersonen Messwerte zugeordnet, zum Beispiel, ein Messwert für das Geschlecht oder die Körpergröße. An diese Zuweisung von Messwerten zu Objekten müssen gewisse Anforderungen gestellt werden, damit sie als Messung aufgefasst werden kann. Die Messwerte wiederum sind nicht zwangsläufig eindeutig. Das *Eindeutigkeitsproblem* in Abschnitt 2.3 zeigt, dass auch Transformationen der Messwerte erlaubt sein können, wie die Messung der Körpergröße in Metern statt in Zentimetern. Das *Bedeutsamkeitsproblem* wirft in Gliederungspunkt 2.4 die Frage auf, welche numerischen Aussagen über die Messwerte möglich sind. So ist zum Beispiel die Berechnung des arithmetischen Mittels für die Haarfarbe der Versuchspersonen nicht sinnvoll. Abschließend muss dann mit Hilfe des *Skalierungsproblems* geklärt werden, wie eine Messung konkret vorgenommen werden kann (2.5).

**Kapitel 3** geht gezielt auf die Frage ein, wie etwas im wirtschaftswissenschaftlichen Kontext gemessen werden kann und wie darüber hinaus die Qualität der Messung beurteilt werden kann. Dafür werden in Punkt 3.1 zunächst verschiedene *Erhebungsmethoden* vorgestellt. Ein Überblick über die Literatur zeigt, dass vor allem die Befragung mit ihren unterschiedlichen Facetten in den Wirtschafts- und Sozialwissenschaften relevant ist. Eine besondere Herausforderung für den Messenden stellt die Tatsache dar, dass interessierende Größen häufig nicht direkt erfassbar sind. Diese Größen, zum Beispiel das Image eines Unternehmens, werden Konstrukte genannt. Kapitel 3.2 stellt die Komplexität der *Konstruktmessung* dar. Im Verlauf der Zeit hat sich eine Vielzahl unterschiedlicher Methoden zur Messung dieser Konstrukte etabliert. Diese verschiedenen *Skalierungsverfahren* werden in Abschnitt 3.3 vorgestellt. Nachdem ein Konstrukt gemessen wurde, muss geklärt werden, wie gut diese Messung gelungen ist. Dazu sind Methoden zur Beurteilung der *Qualität der Messergebnisse* erforderlich (3.4).

**Kapitel 4** thematisiert die Ursachen, die für die schlechte Qualität von Messergebnissen verantwortlich sein können. Als Grundlage wird dafür in Gliederungspunkt 4.1 aufgezeigt, dass die Versuchspersonen bei Befragungen *psychologische Prozesse* durchlaufen, die zu einer Verzerrung der Ergebnisse führen können. In der Literatur existieren viele unterschiedliche Ansätze, die sich mit diesen psychologischen Abläufen befassen. Ein Überblick über die bekanntesten Modelle in den Kapiteln 4.2 bis 4.5 zeigt, dass einige zentrale Gesichtspunkte, etwa Verständnis der Frage, Abrufen und Beurteilung relevanter Information sowie Beantwortung der Frage nahezu immer berücksichtigt werden.

**Kapitel 5** gibt den Stand der Forschung zur konkreten Gestaltung von Ratingskalen wieder. Dabei werden in Punkt 5.1 zunächst Studien vorgestellt, die sich mit der *Beschriftung der Antwortkategorien* befassen. Kapitel 5.2 stellt die Ergebnisse zur der *Verwendung der Antwortkategorie „Keine Meinung“* vor. Abschnitt 5.3 befasst sich mit dem Forschungsstand zur *Anzahl der Antwortkategorien*. Die ausgewählten Studien zeigen, dass die in Gliederungspunkt 4 dargestellten psychologischen Prozesse nur unzureichend oder gar nicht berücksichtigt wurden. Außerdem wenden die Autoren nicht alle erforderlichen Kriterien zur Beurteilung der Qualität einer Ratingskala an. Insbesondere die Validität der Ergebnisse wird häufig nicht ausreichend gewürdigt.

**Kapitel 6** befasst sich deshalb mit einer eigenen empirischen Studie in der die Qualität von Ratingskalen untersucht wird. Dabei wird in Abschnitt 6.1 der *Aufbau der Studie* erläutert. Als Grundlage werden nochmals Problemstellung und Zielsetzung der Untersuchung vor dem Hintergrund der bisherigen Darstellungen in der Arbeit herausgearbeitet. Es folgt die Erläuterung des Untersuchungsdesigns. Gliederungspunkt 6.2 nimmt die Einordnung und Interpretation des *methodischen Vorgehens* und der *Befragungsergebnisse* vor.

**Kapitel 7** dient der kritischen Reflexion der ausgeführten Überlegungen und leitet aus den Befragungsergebnissen des 6. Kapitels Handlungsempfehlungen für die optimale Anzahl von Antwortkategorien bei der Nutzung von Ratingskalen ab. Darüber hinaus wird ein Ausblick auf mögliche weitere Forschungsgebiete gegeben.

Abbildung 1.1 gibt einen graphischen Überblick über Zielsetzung und Struktur der Arbeit:

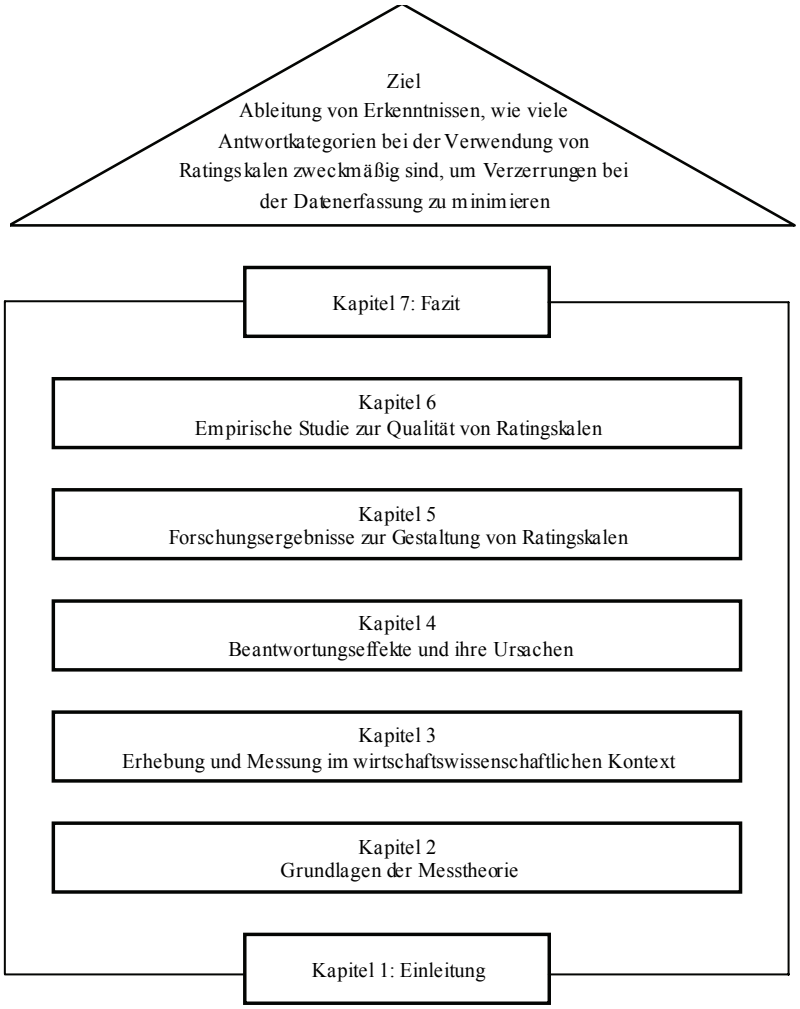


**Abbildung 1.1: Graphische Darstellung der Zielsetzung und des methodischen Aufbaus der Arbeit**

# Kapitel 2

# Grundlagen der Messtheorie

Die Messung von bestimmten Sachverhalten zählt zu den Basiswerkzeugen der Wissenschaft. Damit ist die Durchführung exakter Messungen die Grundlage aller empirischen Wissenschaften, zunächst der Naturwissenschaften, aber auch der Wirtschafts- und Sozialwissenschaften (vgl. ORTH, 1974, S. 9).

Die Messtheorie ist die wissenschaftliche Disziplin, die sich mit der Frage beschäftigt, ob und gegebenenfalls wie physische bzw. psychische Eigenschaften messbar sind. Während sich die Messung naturwissenschaftlicher Phänomene als vergleichsweise einfach erweist, stellt die Messung psychischer Eigenschaften wie Einstellungen, Empfindungen oder Intelligenz auch heute noch eine große Herausforderung dar: „Measurement in the behavioral sciences poses much more difficult and much more heterogeneous problems […]“ (PFANZAGL, 1968, S. 11). Daher liegt der Schwerpunkt dieses Kapitels auf der Messung psychischer Eigenschaften.

Für eine detaillierte Darstellung der zu Grunde liegenden Algebra der Messtheorie sei auf PFANZAGL (1968), KRANTZ, LUCE, SUPPES, TVERSKY (1971), LUCE, KRANTZ, SUPPES, TVERSKY (1990) oder ORTH (1974) verwiesen.

Über Art und Bestandteile der Messung bestehen unterschiedliche Auffassungen. Einige Autoren definieren die Messung als Entwicklung einer geeigneten Messvorschrift (vgl. HAMERLE, 1982, S. 28), andere Autoren, wie beispielsweise STEVENS (1946, S. 677), formulieren dagegen explizit auch die Anwendung einer Messvorschrift als Messung. Beide Ansätze weisen grundlegende Gemeinsamkeiten auf. Die Basis ist stets die Messung der *Ausprägung* einzelner *Merkmale* (Eigenschaften, Variablen, $M_1$, … , $M_m$), die einer Menge von *Objekten*

(Merkmalsträgern, $o_1, \ldots, o_n$) zugeordnet sind. Es ist daher nicht Anspruch der Messtheorie, die Objekte selbst zu messen, sondern stets nur die Ausprägungen der Eigenschaften (vgl. ORTH, 1974, S. 13).

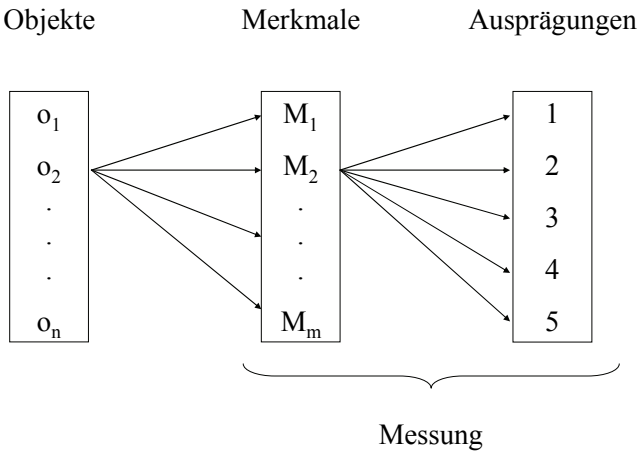


**Abbildung 2.1: Messung im Sinne der Messtheorie**

Die Messung erfolgt, indem den Merkmalsausprägungen der Objekte bestimmte passende Messwerte (Skalenwerte) zugewiesen werden (vgl. Abbildung 2.1).

Die wichtigsten Fragestellungen der Messtheorie können damit wie folgt beschrieben werden (vgl. HILBERT, 1998, S. 7):

1. Ist es für eine gegebene Menge von Objekten $O = \{o_1, \ldots, o_n\}$ möglich, die zwischen den Objekten bestehenden Beziehungen $R_i$ adäquat durch eine Zahlenmenge und darauf definierten Beziehungen $S_i$ so darzustellen, dass jede Beziehung zwischen den Zahlen durch eine empirische Beziehung zwischen den Objekten begründet ist?

2. Wie müssen diese Zahlen aussehen und welche Variationsmöglichkeiten bestehen bei der Wahl der Zahlen, d.h., wie eindeutig sind die Zahlen, welche die Merkmalsausprägungen der Objekte repräsentieren?

Die zur Beantwortung dieser Fragestellungen erforderlichen Schritte können in vier Aufgabengebiete aufgeteilt werden (vgl. ORTH, 1974, S. 39-42):

- Im ersten Schritt muss geklärt werden, welche Bedingungen eine empirische Menge von Objekten erfüllen muss, damit die Zuordnung von

Zahlenwerten sinnvoll ist. Dazu muss jeder Beziehung in der empirischen Menge eine entsprechende Beziehung in der numerischen Menge gegenüberstehen. Dieser Teilbereich wird *Repräsentationsproblem* (Kapitel 2.2) genannt.

- Wenn die Zuordnung von Zahlen sinnvoll erscheint, muss anschließend geklärt werden, wie viel Freiheit bei der Zuordnung von Zahlen zu den Objekten besteht. Dabei steht vor allem die Frage im Mittelpunkt, welche Transformationen der Zahlen zu äquivalenten Ergebnissen führen. Bei dieser Problemstellung spricht man vom *Eindeutigkeitsproblem* (Kapitel 2.3).

- Direkt im Anschluss an das Eindeutigkeitsproblem tritt das *Bedeutsamkeitsproblem* (Kapitel 2.4) auf. Nachdem geklärt ist, welche Transformationen der Zahlen zulässig sind, muss darüber entschieden werden, welche numerischen Aussagen empirisch bedeutsam sind. Es stellt sich z.B. die Frage, ob die Aussage „Objekt $o_1$ ist doppelt so groß wie Objekt $o_2$“ sinnvoll ist.

- Beim vierten zentralen Problem, dem *Skalierungsproblem* (Kapitel 2.5), steht die Frage im Mittelpunkt, wie die konkrete Zuordnung von Zahlen zu Objekten erfolgen soll, wenn die Messbarkeit des zu untersuchenden Merkmals gesichert ist.

Bevor die einzelnen Teilbereiche ausführlicher betrachtet werden, ist es sinnvoll, einige grundlegende Begriffe und Problemstellungen der Messtheorie zu erörtern (vgl. hierzu HILBERT, 1998, S. 8ff).

## 2.1 Begriffe der Messtheorie

Das Ziel der Messtheorie ist es, den Objekten $O = \{o_1, \ldots, o_n\}$, auf Basis der empirisch beobachtbaren oder herstellbaren Beziehungen zwischen den einzelnen Elementen, numerische Werte zuzuordnen. Diese empirisch beobachtbaren Beziehungen können beispielsweise wie folgt aussehen: ´Objekt $o_1$ ist größer als Objekt $o_2$´ oder ´Objekt $o_1$ hat eine andere Haarfarbe als Objekt $o_2$´. Die mathematische Darstellung dieser Beziehungen erfolgt mit Hilfe so genannter Re-

lationen $R_i$ ($i \in I$, I eine endliche Indexmenge). Eine Teilmenge des n-fachen kartesischen Produkts $O \times \ldots \times O$ der Menge $O$ heißt n-stellige Relation $R_i$ auf $O$.

**Definition 2.1**

*Es sei O eine Menge von Objekten und $R_i$ ($i \in I$) seien auf O definierte $r_i$-stellige Relationen. Das $\Theta := \langle O, (R_i)_{i \in I} \rangle$ heißt dann* ***Relativ*** *(vgl. HILBERT, 1998, S. 9).*

In Abhängigkeit der zu Grunde liegenden Menge und der Relationen kann ein numerisches und ein empirisches Relativ unterschieden werden. Bei einem *empirischen Relativ* sind die Elemente der Menge $O$ Personen oder Gegenstände und die Relationen $R_i$ empirisch beobachtbare Relationen. Bei einem *numerischen Relativ* besteht die Menge $O$ aus Zahlen oder Vektoren mit entsprechenden Relationen $R_i$ auf dieser Menge (vgl. PFANZAGL, 1968, S. 19). Zur besseren Unterscheidung soll dafür folgende Bezeichnung gewählt werden: $\Xi := \langle P, (S_i)_{i \in I} \rangle$

Mit diesen Definitionen ist es möglich, eine Messung genauer zu beschreiben. Das empirische Relativ $\Theta$ ist ein Spiegelbild der realen Welt und lässt Rückschlüsse auf diese zu. Bei einer Messung wird dem empirischen Relativ $\Theta$ ein numerisches Relativ $\Xi$ so zugeordnet, dass das numerische Relativ $\Xi$ die Beziehungen des empirischen Relativs $\Theta$ repräsentiert und somit Aussagen über das empirische Relativ und die reale Welt ermöglicht (vgl. Abbildung 2.2). Dies bedeutet beispielsweise, dass der empirischen Beziehung ´Objekt $o_1$ ist größer als Objekt $o_2$´ auch im numerischen Relativ $\Xi$ durch ´$p_1 > p_2$´ Rechnung getragen wird. Auf welche Weise das numerische Relativ dem empirischen Relativ zugeordnet wird, soll im Folgenden geklärt werden.

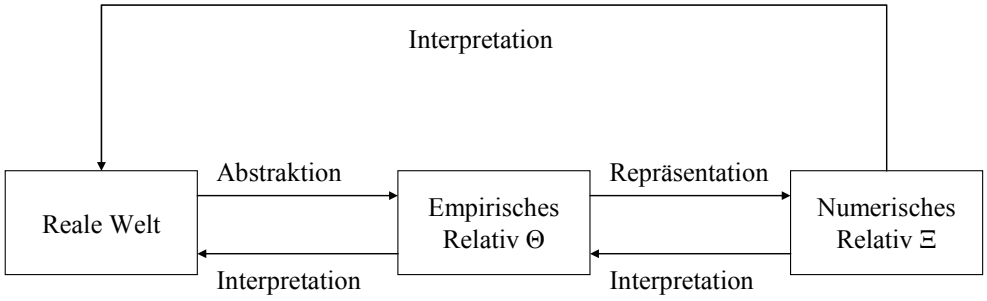


**Abbildung 2.2: Schematische Darstellung der Repräsentationsbeziehungen (vgl. HILBERT, 1998, S. 9)**

Damit das numerische Relativ das empirische Relativ repräsentiert, sollten alle Relationen $R_i$ auf den Objekten des empirischen Relativs $\Theta$ durch entsprechende Relationen $S_i$ auf dem numerischen Relativ $\Xi$ abgebildet werden. Die folgenden Darstellungen beziehen sich (o.B.d.A.) auf Relative mit ausschließlich 2-stelligen Relationen.

**Definition 2.2**

*Es seien $\Theta := \langle O, R_1, \ldots, R_n \rangle$ und $\Xi := \langle P, S_1, \ldots, S_n \rangle$ zwei Relative. Eine Abbildung f von O in P heißt* ***Homomorphismus*** *(oder homomorphe Abbildung) von O in P, wenn für alle Elementpaare $o_1, o_2 \in O$ und für jede Relation $R_i$ gilt:*

$$(o_1, o_2) \in R_i \Leftrightarrow (f(o_1), f(o_2)) \in S_i$$

Liegt eine homomorphe Abbildung des empirischen Relativs $\Theta$ in ein numerisches Relativ $\Xi$ vor, so spricht man auch von einer **Skala**. Das Bild eines Objektes $o \in O$ heißt dann Messwert oder Skalenwert von $o$.

Mit den bisher getroffenen Definitionen kann abschließend der Begriff *Messung* definiert werden.

**Definition 2.3**

*„**Messen** ist die Bestimmung der Ausprägung einer Eigenschaft eines Dinges. Messen erfolgt durch eine Zuordnung von numerischen Größen (Zahlen, Vektoren) zu Dingen, die Träger der zu messenden Eigenschaft sind. Messen beruht auf einer homomorphen Abbildung eines empirischen Relativs in ein numerisches Relativ bzw. auf einer Repräsentation eines empirischen Relativs durch ein numerisches Relativ. Die Existenz einer derartigen homomorphen Abbildung ist das Kriterium dafür, ob eine Zuordnung von Zahlen zu Dingen als ‚Messen' zu betrachten ist, d.h. ob eine Eigenschaft messbar ist"* (ORTH, 1974, S. 18).

Diese Definition beinhaltet keine Aussage darüber, ob alle Relationen des empirischen Relativs $\Theta$ bei der Konstruktion einer Skala zu berücksichtigen sind. Da bei einer Messung so viel Information wie möglich dargestellt werden soll, erscheint es sinnvoll, durch eine Skala alle bekannten empirischen Relationen abzubilden und sich nicht nur auf eine Teilmenge zu beschränken (vgl. PFANZAGL, 1968, S. 26-27). Es bleibt festzuhalten, dass eine Messung ein

Homomorphismus eines empirischen Relativs in ein numerisches Relativ ist, bei dem möglichst alle Relationen des empirischen Relativs geeignet übertragen werden. Im nächsten Schritt ist zu klären, wann ein empirisches Relativ als messbar gilt, d.h. welche Voraussetzungen ein empirisches Relativ erfüllen muss, damit es durch ein numerisches Relativ repräsentiert werden kann.

## 2.2 Das Repräsentationsproblem

Falls ein Homomorphismus des empirischen Relativs in ein numerisches Relativ existiert, ist die Eigenschaft von Objekten messbar. Bei der Messung der Länge von Objekten liegen im empirischen Relativ $\Theta := \langle O, (R_i)_{i \in I} \rangle$ beispielsweise die Relationen ´$\prec$´ (kleiner als), ´$\sim$´ (gleich groß wie) und ´$o$´ (Aneinanderlegen der Objekte) vor (vgl. ORTH, 1974, S. 21). Das numerische Relativ $\Xi := \langle P, (S_i)_{i \in I} \rangle$ besteht analog aus den Relationen ´$<$´, ´$=$´ und ´$+$´. Die Repräsentation muss dann folgende Bedingungen erfüllen:

- Wenn ein Objekt länger als ein anderes Objekt ist, wird ihm ein größerer Skalenwert zugewiesen.
- Der Skalenwert, der zwei aneinander gelegten Objekten zugewiesen wird, muss der Summe der Einzelskalenwerte entsprechen.

Wenn beide Bedingungen erfüllt sind, ist das Repräsentationsproblem gelöst und die Eigenschaft messbar.

**Beispiel 2.1**

*Für eine Menge O mit 5 Objekten sei bezogen auf das Merkmal ´Größe´ folgende transitive Relation bekannt:*

$$o_1 \prec o_2 \prec o_3 \prec o_4 \sim o_5$$

*Zwei mögliche Homomorphismen des empirischen Relativs $\Theta := \langle O, \prec, \sim \rangle$ in das numerische Relativ $\Xi := \langle P, <, = \rangle$ sind in Abbildung 2.3 dargestellt:*

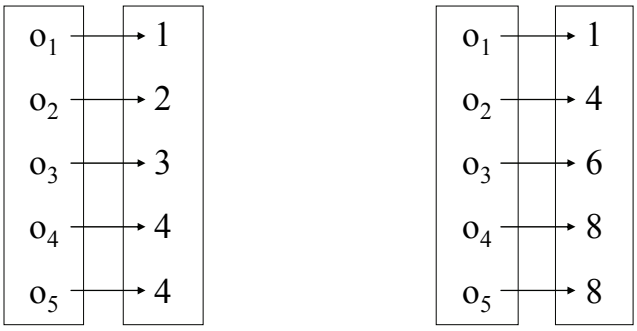


**Abbildung 2.3: Beispiel für Homomorphismen**

Es kann also im Allgemeinen mehr als eine geeignete homomorphe Abbildung eines empirischen Relativs $\Theta$ in ein numerisches Relativ $\Xi$ existieren. Die Menge aller Homomorphismen sei mit $HOM(\Theta,\Xi)$ bezeichnet. Die Frage, welche Transformationen der Skalenwerte zulässig sind, entspricht dem Eindeutigkeitsproblem der Messung. Erweitert man in Beispiel 2.1 die zur Verfügung stehende Information dahingehend, dass $o_1$ und $o_2$ aneinandergelegt genauso groß sind wie $o_3$, dann stellt nur noch der linke Teil in Abbildung 2.3 einen Homomorphismus dar. Die Transformation auf der rechten Seite wäre nicht zulässig.

## 2.3 Das Eindeutigkeitsproblem

In der Regel existieren zu einem Repräsentationsproblem mehrere Skalen, die einen Homomorphismus bilden (vgl. HAMERLE, 1982, S. 32). Die Menge HOM(Θ,Ξ) besteht also aus mehr als einem Element. Im ersten Schritt zur Lösung des Eindeutigkeitsproblems wird geklärt, ob zwei homomorphe Abbildungen aus HOM(Θ,Ξ) durch eine Transformation γ ineinander überführt werden können. Eine Transformation γ heißt *zulässig*, wenn die Eigenschaft das Repräsentationsproblem zu lösen erhalten bleibt (vgl. ORTH, 1974, S. 24). Im zweiten Schritt werden alle zulässigen Transformationen γ zu einer Menge Γ zusammengefasst. Die Menge Γ bestimmt dann den Eindeutigkeitsgrad einer Skala. Fasst man im nächsten Schritt alle Skalen mit einer identischen Menge Γ

an zulässigen Transformationen zusammen, so kann dadurch schließlich der Skalentyp oder die Skalenart bestimmt werden (vgl. HILBERT, 1998, S. 12).

Die Identifizierung der zulässigen Transformationen und damit die Festlegung des Skalentyps, stellt somit die Lösung des Eindeutigkeitsproblems dar. Nach STEVENS (1946, 1951) können folgende Skalentypen und zulässige Transformationen unterschieden werden:

| **Skalentyp** | **Zulässige Transformationen** | **Invariant** |
|---|---|---|
| Nominalskala | Injektive Funktionen | Eindeutigkeit der Messwerte |
| Ordinalskala | Monoton steigende Funktionen | Rangordnung der Messwerte |
| Intervallskala | positiv-lineare Funktionen: $\gamma(x) = ax + b, (a,b) \in \Re, a > 0$ | Verhältnisse der Intervalle zwischen den Messwerten |
| Verhältnisskala | Ähnlichkeitsfunktion: $\gamma(x) = ax, (a > 0)$ | Verhältnisse der Messwerte |
| Absolutskala | $\gamma(x) = x$ | Messwerte |

**Tabelle 2.1: Skalentypen und zulässige Transformationen (in Anlehnung an ORTH, 1974, S. 24)**

Die *Nominalskala* besitzt das geringste Skalenniveau, die Menge $\Gamma$ umfasst in diesem Fall alle injektiven Funktionen. Die Zuordnung von Zahlen zu den einzelnen Merkmalsausprägungen ist zufällig. Bei der Interpretation erfolgt lediglich die Überprüfung, ob die Skalenwerte zweier Objekte übereinstimmen oder nicht. Somit ergibt sich für Merkmale mit nominalem Skalenniveau die Forderung, dass übereinstimmende Merkmalsausprägungen zum selben Skalenwert führen und unterschiedliche Ausprägungen verschiedene Skalenwerte ergeben.

$$o_1 = o_2 \quad \Rightarrow \quad f(o_1) = f(o_2) \qquad\qquad o_1 \neq o_2 \quad \Rightarrow \quad f(o_1) \neq f(o_2)$$

Jede injektive Abbildung kann den Informationsgehalt nominaler Variablen darstellen und ist daher zulässig. Somit handelt es sich bei einer Nominalskala lediglich um die Zuordnung von Messwerten zu Merkmalsausprägungen, wie bei Haarfarbe oder Geschlecht einer Person.

Bei der *Ordinalskala* sinkt die Zahl der zulässigen Transformationen im Vergleich zur Nominalskala. Eine zulässige Transformation muss die Rangordnung der Skalenwerte beibehalten. Daher sind für ordinale Merkmale alle monoton steigenden Transformationen zulässig (vgl. ALLEN, YEN, 1979, S. 7).

$$o_1 \prec o_2 \quad \Rightarrow \quad f(o_1) < f(o_2) \qquad o_1 \succ o_2 \quad \Rightarrow \quad f(o_1) > f(o_2)$$

Bei der hier getroffenen Definition liegen Ordinalskalen ausschließlich bei streng monoton steigenden Abbildungen vor (vgl. PFANZAGL, 1968, S. 76). Typische Beispiele ordinaler Merkmale, etwa Schulnoten, bilden nach dieser Definition eigentlich keine Ordinalskala, da Objekten mit leicht unterschiedlichen Leistungen keine identischen Noten zugeordnet werden dürften. Dies führt zu einer etwas allgemeineren Definition:

$$o_1 \prec o_2 \quad \Rightarrow \quad f(o_1) \leq f(o_2) \qquad o_1 \succ o_2 \quad \Rightarrow \quad f(o_1) \geq f(o_2)$$

Für die Ordinalskala sind Aussagen über die Rangfolge der Objekte erlaubt. Eine Beurteilung der Distanz zwischen zwei Merkmalsausprägungen ist jedoch nicht möglich. Für die Interpretation der Distanz zwischen zwei Merkmalsausprägungen ist mindestens eine *Intervallskala* erforderlich. Bei ihr sind alle positiv linearen Transformationen zulässig, „die nicht nur die Rangordnung der Meßwerte, sondern auch die Verhältnisse von Intervallen (Differenzen) zwischen Meßwerten invariant lassen“ (ORTH, 1974, S. 25).

$$o_1 - o_2 \prec o_3 - o_4 \quad \Rightarrow \quad f(o_1) - f(o_2) < f(o_3) - f(o_4)$$

Eine typische Intervallskala ist die Messung der Temperatur in Celsius oder Fahrenheit. Bei einer Intervallskala sind zwar Aussagen über die Differenz der Merkmalsausprägungen zulässig, nicht aber über das Verhältnis der Messwerte (vgl. KLEIN, SCHOLL, 2004, S. 33). Eine scheinbare Verdoppelung der Temperatur gemessen in Celsius entspricht nicht einer Verdoppelung gemessen in Fahrenheit. Falls nicht nur die Verhältnisse der Differenzen zwischen Messwerten invariant bleiben, sondern darüber hinaus auch die Verhältnisse der Messwerte an sich, spricht man von einer *Verhältnisskala*.

$o_1 : o_2 \prec o_3 : o_4 \quad \Rightarrow \quad f(o_1) : f(o_2) < f(o_3) : f(o_4)$

Körpergröße, Gewicht aber auch die absolute Temperatur gemessen nach Calvin sind typische Vertreter der Verhältnisskala. Ist bei der Messung der Körpergröße eine Person doppelt so groß wie eine andere, so ändert sich dieses Verhältnis nicht, wenn statt in Zentimetern in Metern gemessen wird. Dies liegt an der Existenz eines natürlichen Nullpunktes, der bei der Intervallskala fehlt. Die Einheit, in der bei einer Verhältnisskala gemessen wird, ist jedoch willkürlich. So kann die Körpergröße in Metern, Zentimetern aber auch Dezimetern gemessen werden.

Den größten Informationsgehalt liefert die *Absolutskala*, wie zum Beispiel die Anzahl der Geschwister. Mit Ausnahme der Identitätstransformation sind keine weiteren Transformationen möglich.

Eine nähere Betrachtung ist für die Unterscheidung zwischen Ordinalskala und Intervallskala erforderlich, da diese in der Literatur kontrovers diskutiert wird. Es lässt sich nämlich zeigen, dass unter bestimmten Voraussetzungen die Ordinalskala intervallskaliert ist. Dazu müssen die Skalenwerte der Ordinalskala äquidistant sein.

**Bemerkung 2.1**

*Ein Merkmal mit den Ausprägungen $a_0 < a_1 < \ldots < a_k$ ist intervallskaliert*

$\Leftrightarrow \quad f(a_i) = b_i = \alpha a_i + \beta \; (\alpha > 0, \beta \in \Re)$.

***Satz 1***

*Für ein ordinales Merkmal sei $(a_{i+1} - a_i) = c$ $(i = 0, \ldots, k\text{-}1)$. Dann ist das Merkmal intervallskaliert.*

***Beweis***

*Sei $a_i = a_0 + ic$, $b_i = b_0 + id$ $(i = 0, 1, \ldots, k\text{-}1)$, $c > 0$, $d > 0$*

$$\Rightarrow i = \frac{a_i - a_0}{c}$$

$$b_i = b_0 + \frac{a_i - a_0}{c} d = \frac{d}{c} a_i + b_0 - \frac{d}{c} a_0 = \alpha a_i + \beta$$

$$\text{für } \alpha = \frac{d}{c}, \quad \beta = b_0 - \frac{d}{c} a_0$$

*Die Umkehrung gilt nicht allgemein.*

### *Satz 2*

*Für ein intervallskaliertes Merkmal mit $b_i = \alpha a_i + \beta$ sei $(a_{i+1} - a_i) = c$. Dann ist $(b_{i+1} - b_i)$ konstant für alle i.*

### *Beweis*

$$b_{i+1} - b_i = \alpha a_{i+1} + \beta - \alpha a_i - \beta = \alpha(a_{i+1} - a_i) = c$$

Wie Bemerkung 2.1 zeigt, ist die Behandlung der Ordinalskala als intervallskaliertes Merkmal gerechtfertigt, wenn die Messwerte der Ordinalskala äquidistant sind. Dennoch kann die Diskussion über die Behandlung ordinaler Daten in der Literatur weit zurückverfolgt werden. So unterscheiden bereits ACOCK und MARTIN (1974, S. 427) zwischen methodischen Puristen und methodischen Pragmatikern. Die methodischen Pragmatiker verstehen ordinale Merkmale häufig lediglich als Konstrukt, um intervallskalierte Sachverhalte zu messen, zum Beispiel den (ordinalen) Intelligenzquotienten zur Messung der (intervallskalierten) Intelligenz (vgl. LABOVITZ, 1967, S. 152). Diese Sichtweise wird zusätzlich dadurch gestützt, dass in der realen Datenerhebung häufig intervall- oder gar verhältnisskalierte Indikatoren zur Messung von ordinalen Merkmalen herangezogen werden. So erfolgt die Messung des Intelligenzquotienten unter anderem durch die Zeitdauer, die eine Person für eine bestimmte Aufgabenstellung benötigt. Die Pragmatiker argumentieren, dass diese intervallskalierten Indikatoren eine Näherung der unbekannten intervallskalierten Sachverhalte darstellen (vgl. LABOVITZ, 1967, S. 153). Für sie liegt der Schluss daher nahe, diese Variablen als intervallskaliert anzusehen. Die methodischen Puristen betonen dagegen, dass auch in diesem Fall die resultierende Skala nur ordinales Datenniveau besitzt, da die genauen Zusammenhänge zwischen der Zeitdauer und der Intelligenz unbekannt sind (vgl. PFANZAGL, 1968, S. 77).

Eine andere Einteilung der verschiedenen Skalentypen soll hier abschließend kurz erwähnt werden. Für nominale und ordinale Merkmale ist auch der Begriff *Topologische Skala* oder *qualitatives Merkmal* üblich. Intervall-, Verhältnis- und Absolutskala werden dagegen als *Kardinalskala* oder *quantitatives Merkmal* bezeichnet (vgl. HILBERT, 1998, S. 14).

Nachdem durch die Lösung des Eindeutigkeitsproblems der Skalentyp einer Messung bekannt ist, muss das jeweilige Skalenniveau in der späteren

Datenanalyse problemadäquat behandelt werden. Dies impliziert die Lösung des Bedeutsamkeitsproblems.

## 2.4 Das Bedeutsamkeitsproblem

Im Allgemeinen bilden die Messwerte an sich nur die Basis, um darauf aufbauend weitere Werte zu berechnen (vgl. ORTH, 1974, S. 29). Dabei sind vor allem Lageparameter (z.B. arithmetisches Mittel), Streuungsparameter (z.B. Varianz) und Zusammenhangsmaße (z.B. Bravais-Pearson-Korrelationskoeffizient) relevant. Allerdings können nicht alle statistischen Kennzahlen für jedes Skalenniveau angewandt werden. Das Bedeutsamkeitsproblem beantwortet die Frage, welche Kennzahlen und Maße für ein Merkmal geeignet sind. Damit stellt sich das Bedeutsamkeitsproblem erst bei der Datenauswertung und nicht schon bei der Messung selbst.

Berechnet man für eine Objektmenge den Mittelwert der Körpergröße, so ist dies eine *numerische Aussage*. Falls der Wahrheitswert der numerischen Aussage bei allen zulässigen Transformationen der zu Grunde gelegten Skala erhalten bleibt, heißt sie *bedeutsam* (PFANZAGL, 1968, S.34-36). Das Bedeutsamkeitsproblem ist somit die Feststellung, ob eine numerische Aussage für die vorhandene Skala sinnvoll ist (vgl. ORTH, 1974, S. 30).

**Beispiel 2.2**

*Die Beurteilung einer Klausur erfolgt mit Hilfe von Bewertungspunkten. Diese Bewertungspunkte werden in zwei verschiedene Notenskalen überführt, die Schulnotenskala mit Noten von 1-6 und die Kollegstufenskala mit einer Einteilung von 0-15 Punkten. Beide Einteilungen sind zulässige Transformationen der erreichten Bewertungspunkte in eine Notenskala. Die Möglichkeit dieser Einteilung entspricht der Lösung des Repräsentationsproblems. Dabei gelte folgender Punkteschlüssel:*

| Bewertungspunkte | Schulnotenskala | Kollegstufenskala |
|---|---|---|
| 50,49 | 1 | 15 |
| 48,47 | 1 | 14 |
| 46,45 | 1 | 13 |
| 44,43 | 2 | 12 |
| 42,41 | 2 | 11 |
| 40,39 | 2 | 10 |
| 38,37 | 3 | 9 |
| 36,35 | 3 | 8 |
| 34,33 | 3 | 7 |
| 32,31 | 4 | 6 |
| 30,29 | 4 | 5 |
| 28,27 | 4 | 4 |
| 26,25 | 5 | 3 |
| 24,23 | 5 | 2 |
| 22,21 | 5 | 1 |
| 20-0 | 6 | 0 |

**Tabelle 2.2: Bewertung einer Klausur in Abhängigkeit der erreichten Punkte**

*Für die fünf Schüler A, B, C, D und E sind die erreichten Punktzahlen einer Klausur bekannt.*

| | Bewertungspunkte | Schulnotenskala | Kollegstufenskala |
|---|---|---|---|
| A | 50 | 1 | 15 |
| B | 49 | 1 | 15 |
| C | 41 | 2 | 11 |
| D | 37 | 3 | 9 |
| E | 31 | 4 | 6 |

**Tabelle 2.3: Transformation der Punktzahlen**

*Ein Vergleich der Objektpaare (B,C) und (C,D) zeigt, dass beide Noteneinteilungen nicht intervallskaliert sind. Bei Verwendung der Schulnoten ist die Distanz der beiden Objektpaare identisch. Nach Durchführung der ebenfalls zulässigen Transformation in die Kollegstufenpunkte ist die Distanz zwischen B und C mit 4 deutlich größer, als zwischen C und D mit 2. Daher handelt es sich bei Schulnoten und Kollegstufenskala um eine Ordinalskala. Die Lösung des Eindeutigkeitsproblems liefert somit das Skalenniveau der Messung.*

*Berechnet man anschließend das arithmetische Mittel und den Median für beide Noteneinteilungen, so erhält man folgendes Ergebnis:*

| | Arithmetisches Mittel | Median |
|---|---|---|
| Schulnote | 2,2 | 2 |
| Kollegstufenpunkte | 11,2 | 11 |

**Tabelle 2.4: Lageparameter der Schulnoten**

*Während der Schüler C bei der Einteilung nach Schulnoten besser als das arithmetische Mittel abschneidet, ist er bei Verwendung der Kollegstufenpunkte schlechter als das arithmetische Mittel. Die Verwendung des arithmetischen Mittels ist demnach für ordinale Daten nicht sinnvoll. Anders verhält es sich beim Median. Sowohl bei der Einteilung nach Schulnoten, als auch bei den Kollegstufenpunkten entspricht die Note des Schülers C exakt dem Median. Der Median ist demnach für ordinale Merkmale geeignet. Die Lösung des Bedeutsamkeitsproblems liefert somit die anwendbaren statistischen Kennzahlen.*

In den nachfolgenden Tabellen sollen die zulässigen statistischen Verfahren für die jeweiligen Skalenniveaus dargestellt werden, ohne explizit an dieser Stelle schon auf die Maßzahlen einzugehen.

| | Modus | Median | Arithmetisches Mittel | Geometrisches Mittel |
|---|---|---|---|---|
| Nominalskala | Ja | - | - | - |
| Ordinalskala | Ja | Ja | - | - |
| Intervallskala | Ja | Ja | Ja | - |
| Verhältnisskala | Ja | Ja | Ja | Ja |

**Tabelle 2.5: Zulässige Lageparameter**

In Tabelle 2.5 sind die möglichen Lageparameter für die einzelnen Skalenniveaus dargestellt. Prinzipiell kann davon ausgegangen werden, dass für Skalen mit höherem Informationsgehalt alle Maße erlaubt sind, die auch für niedriger skalierte Merkmale zulässig sind. Ähnlich verhält es sich auch bei den Streuungsparametern (siehe Tabelle 2.6):

| | Modal-abweichung | Perzentile | Varianz | Variations-koeffizient |
|---|---|---|---|---|
| Nominalskala | Ja | - | - | - |
| Ordinalskala | Ja | Ja | - | - |
| Intervallskala | Ja | Ja | Ja | - |
| Verhältnisskala | Ja | Ja | Ja | Ja |

**Tabelle 2.6: Zulässige Streuungsparameter**

Aus messtheoretischer Sicht sind damit die Grundlagen für eine Messung gelegt. Es bleibt jedoch die Frage bestehen, auf welche Art und Weise den Objekten konkrete Zahlen und Messwerte zugeordnet werden können. Da dieses Problem in der Praxis von äußerst großer Relevanz ist, darf auch in der Messtheorie auf eine nähere Betrachtung nicht verzichtet werden. Die Messtheorie befasst sich mit dieser Fragestellung im Rahmen des Skalierungsproblems.

## 2.5 Das Skalierungsproblem

In der Messtheorie steht die Frage nach der Messbarkeit von Eigenschaften im Vordergrund. Daher behandeln die in den Kapiteln 2.2 bis 2.4 vorgestellten Problembereiche die Frage, ob die Messung einer Eigenschaft möglich ist. Das Skalierungsproblem erweitert diese Fragestellung dahingehend, dass zusätzlich die Frage, wie eine Eigenschaft gemessen werden kann, in den Fokus rückt. Dabei bestehen sehr viele unterschiedliche Möglichkeiten zur konkreten Vorgehensweise bei einer Messung, die im Rahmen dieses Kapitels nicht vorgestellt werden können. Die bedeutsamsten Skalierungsverfahren werden in Kapitel 3 dieser Arbeit behandelt. An dieser Stelle sollen lediglich einige grundlegende Eigenschaften der Skalierungsverfahren diskutiert werden.

Grundsätzlich gilt für alle Skalierungsverfahren, dass die Möglichkeit der Zuordnung von Messwerten zu den Objekten vom Merkmal selbst abhängig ist (vgl. HILBERT, 1998, S. 16). Dabei ist vor allem die Anzahl möglicher Ausprägungen von Bedeutung. Bei einigen Merkmalen existieren lediglich abzählbar viele Ausprägungen, etwa die Anzahl der Geschwister. Für diese Merkmale hat sich in der Literatur der Begriff *diskret* durchgesetzt. Andere Merkmale haben dagegen theoretisch überabzählbar viele Ausprägungen. Diese werden als *stetig* bezeichnet. Die Körpergröße ist ein typisches Beispiel eines stetigen Merkmals. Eine Sonderrolle nehmen die so genannten *quasi-stetigen* Merkmale ein. Quasi-stetige Merkmale können zum einen ursprünglich diskret sein, aber auf Grund ihrer Vielzahl an Ausprägungen erscheint es zweckmäßig, sie als stetig zu behandeln. Zum anderen kann auf Grund fehlender Messgenauigkeit ein stetiges Merkmal diskret erscheinen (vgl. BAMBERG, BAUR, 2002, S. 7). Beispielsweise existieren bei der Körpergröße gemessen in

Zentimetern nicht unendlich viele Ausprägungen. Dies liegt aber nicht in der Natur des Merkmals, sondern an der ungenauen Durchführung der Messung.

Es ist zu beachten, dass die Einteilung in stetige und diskrete Merkmale unabhängig vom Skalenniveau ist (vgl. HILBERT, 1998, S. 11), wie Tabelle 2.7 zeigt:

| | Diskrete Merkmale | Stetige Merkmale |
|---|---|---|
| Qualitative Merkmale | Geschlecht<br>Nationalität | Farbtöne<br>Intelligenzquotient |
| Quantitative Merkmale | Einwohnerzahl<br>Anzahl Geschwister | Größe<br>Alter |

**Tabelle 2.7: Beispiele für stetige und diskrete Merkmale in Abhängigkeit des Skalenniveaus**

Skalierungsverfahren können streng genommen nur dann eingesetzt werden, wenn die untersuchte Eigenschaft messbar ist. Es existieren allerdings auch Skalierungsverfahren, die die Messbarkeit der Eigenschaft nicht überprüfen (vgl. ORTH, 1974, S. 41). Wenn die Messbarkeit einer Eigenschaft nicht überprüft wird, liefern Skalierungsverfahren keine mathematisch exakt begründeten Aussagen über die Messwerte (vgl. HILBERT, 1998, S. 11). Insbesondere in komplexen Themengebieten, etwa der Messung der Intelligenz, kann die Frage nach der Messbarkeit und der geeigneten Konstruktion eines Homomorphismus jedoch kaum beantwortet werden. In solchen Fällen stellen Skalierungstechniken ohne messtheoretische Grundlagen die einzige Möglichkeit dar, den betreffenden Sachverhalt näher zu untersuchen. In diesem Zusammenhang wird von einem Messverfahren ´per fiat´ gesprochen, da es auf dem Glauben beruht, dass die betreffende Eigenschaft auch messbar ist (vgl. ORTH, 1974, S. 41).

## 2.6 Zusammenfassung und Einordnung

Dieses Kapitel hat vor allem gezeigt, welche Anforderungen an eine Eigenschaft bestehen, damit diese messbar ist. Dazu ist die Existenz einer homomorphen Abbildung eines empirischen Relativs in ein numerisches Relativ erforderlich. Mit Hilfe der Messtheorie kann somit geklärt werden, ob und unter welchen Voraussetzungen die Messung eines Sachverhalts durchgeführt werden kann.

Darüber hinaus liefert die Messtheorie erste Anhaltspunkte für die Beurteilung der Qualität einer Messung. Ziel der Messung ist es nämlich, möglichst alle Relationen des empirischen Relativs im numerischen Relativ zu erhalten (siehe Seite 10). Insofern bildet dieses Kapitel das theoretische Grundgerüst für die Arbeit.

Allerdings liefert die Messtheorie keine Anhaltspunkte dafür, wie eine Messung im wirtschafts- und sozialwissenschaftlichen Bereich konkret operationalisiert und wie die Qualität anschließend bewertet werden kann. Mit diesen Fragestellungen beschäftigt sich Kapitel 3.

# Kapitel 3

# Erhebung und Messung im wirtschaftswissenschaftlichen Kontext

Nachdem in Kapitel 2 die theoretischen Grundlagen einer Messung besprochen wurden, soll in diesem Kapitel auf konkrete Aspekte bei Erhebungen im wirtschafts- und sozialwissenschaftlichen Umfeld eingegangen werden. Jetzt steht somit die Beantwortung der Frage, wie etwas gemessen werden kann, im Zentrum der Ausführungen.

Dazu sollen zunächst die verschiedenen *Erhebungsmethoden* in diesem Bereich geschildert werden (Kapitel 3.1).

Im Anschluss daran wird auf die *Messung von Konstrukten* näher eingegangen (Kapitel 3.2). In diesem Kapitel wird das im wirtschaftswissenschaftlichen Kontext häufig auftretende Phänomen thematisiert, dass die interessierende Variable nicht direkt beobachtet bzw. gemessen werden kann.

Der Zusammenhang zwischen einer nicht beobachtbaren interessierenden Variable und den messbaren Größen wird durch *Skalierungsverfahren* herausgearbeitet (Kapitel 3.3). Skalierungsverfahren sind somit Messinstrumente, die aufbauend auf der in Kapitel 2 beschriebenen Messtheorie mit Hilfe der Erhebungsmethoden aus Kapitel 3.1 eine konkrete Messung vornehmen.

Im letzten Teil des Kapitels sollen die mit Hilfe der Skalierungsverfahren erhaltenen Messergebnisse auf ihre *Qualität* untersucht werden (Kapitel 3.4).

## 3.1 Erhebungsmethoden

In diesem Kapitel wird diskutiert, auf welchen Wegen im wirtschaftswissenschaftlichen Umfeld die Informationsgewinnung erfolgen kann. Zunächst müssen dazu folgende Erhebungsmethoden voneinander abgegrenzt werden:

- Primärerhebung
- Sekundärerhebung

Bei der *Sekundärerhebung* wird auf bereits existierende Informationsquellen zurückgegriffen, z.B. von Meinungsforschungsinstituten oder dem statistischen Bundesamt (vgl. GIERL, 1995, S. 207). Bei der *Primärerhebung* werden die Informationen dagegen neu ermittelt. Während bei der Primärerhebung die in diesem Kapitel diskutierten Problemstellungen direkt für das Unternehmen bzw. den Untersuchungsleiter anfallen, muss man bei der Sekundärerhebung der Informationsquelle vertrauen.

Vorteile der Sekundärerhebung sind die schnelle Verfügbarkeit der Informationen und die vergleichsweise geringen Kosten. Die Nachteile der Sekundärforschung sind vielfältiger Natur, unter anderem können bei folgenden Aspekten Probleme auftreten (vgl. MEFFERT, 1992, S. 196):

- Aktualität der Daten
- Vergleichbarkeit bei unterschiedlichen Informationsquellen
- Genauigkeit der Information
- Detaillierungsgrad

Falls man auf die Primärerhebung zurückgreift, so bieten sich vor allem zwei Möglichkeiten zur Operationalisierung an (vgl. GIERL, 1995, S. 206):

- Beobachtung (Kapitel 3.1.1)
- Befragung (Kapitel 3.1.2)

Ein weiteres von diesen beiden Methoden nicht klar abgrenzbares, Verfahren zur Informationsgewinnung ist das *Experiment*. Ein Experiment liegt immer dann

vor, wenn das Untersuchungsdesign die Manipulation (Kontrolle) einer Variablen ermöglicht (vgl. KALLMANN, 1979, S. 48). Mit Manipulation einer Variablen ist gemeint, dass der Einfluss dieser Variable auf die zu untersuchende Zielvariable durch den Untersuchungsleiter kontrolliert wird, um Kausalzusammenhänge zwischen interessierenden Merkmalen ohne störende Drittvariable zu messen. Dazu zählt beispielsweise die Konstanthaltung genauso wie die gezielte Variation der Temperatur im Rahmen eines physikalischen Experiments. Ein Merkmal, dessen Einfluss nicht kontrollierbar ist, fließt als Störvariable in das Experiment ein (vgl. MEFFERT, 1992, S. 207). Abbildung 3.1 veranschaulicht die Elemente eines Experiments.

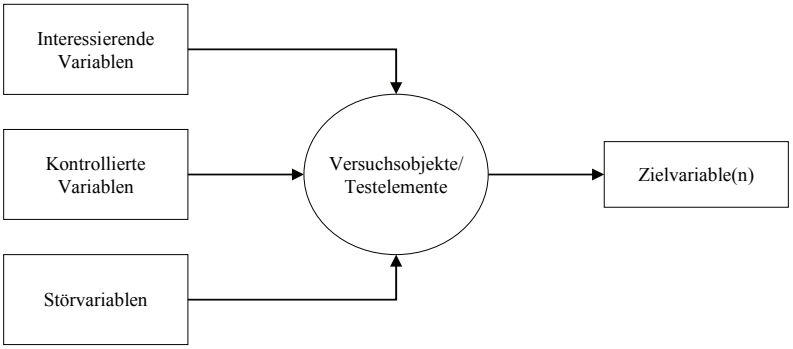


**Abbildung 3.1: Elemente eines Experiments (in Anlehnung an MEFFERT, 1992, S. 208)**

Bei der konkreten Messung greift ein Experiment auf Beobachtungen oder Befragungen zurück, weshalb es keine eigenständige Erhebungsmethode ist (vgl. MEFFERT, 1992, S. 208). Während in den Naturwissenschaften dem Experiment große Bedeutung zukommt, spielt es in den Wirtschaftswissenschaften eher eine untergeordnete Rolle (vgl. HAMMANN, ERICHSON, 2000, S. 158). Dies liegt vor allem an der Variablenmanipulation, die in den Naturwissenschaften deutlich einfacher zu bewerkstelligen ist. So sind Einflussgrößen wie Temperatur, Luftdruck oder Luftfeuchtigkeit durch Konstanthaltung leichter beeinflussbar, als menschliche Wesenszüge, Stimmungslagen oder Produktalternativen (vgl. HAMMANN, ERICHSON, 2000, S. 158). Am ehesten kommen bei psychologischen Messungen daher so genannte Feldexperimente zum Einsatz, die unter realen Bedingungen die Manipulation einiger Variablen ermöglichen (vgl. KALLMANN, 1979, S. 49).

### 3.1.1 Beobachtung

Bei der Beobachtung werden die relevanten Merkmale visuell oder instrumentell erhoben (vgl. HAMMANN, ERICHSON, 2000, S. 98). Daher stellt die Beobachtung nur für objektive Tatbestände, etwa physische Aktivitäten, Verhaltensweisen oder soziodemographische Daten ein sinnvolles Messinstrument dar (vgl. MEFFERT, 1992, S. 198). Für subjektive Sachverhalte sind Beobachtungen nur dann geeignet, wenn die psychischen Phänomene physische Konsequenzen nach sich ziehen, etwa eine Veränderung der Pulsfrequenz oder des Gesichtsausdrucks (vgl. KROEBER-RIEL, WEINBERG, 1999, S. 32ff.).

Von einer *Fremdbeobachtung* spricht man, wenn ein Beobachter Vorgänge dokumentiert, die außerhalb seiner Person liegen (vgl. MEFFERT, 1992, S. 198). Bei der *Selbstbeobachtung* ist die Versuchsperson gleichzeitig der Beobachter. Diese Form der Beobachtung liegt beispielsweise dann vor, wenn die Person eigene psychische Vorgänge berichtet (vgl. ROGGE, 1981, S. 127).

Wird die Auskunftsperson über eine Fremdbeobachtung informiert, so handelt es sich um eine *offene* Beobachtung, andernfalls spricht man von einer *verdeckten* Beobachtung (vgl. HAMMANN, ERICHSON, 2000, S. 98). Gegenüber der verdeckten Beobachtung greift die offene Form nicht in die Persönlichkeitsrechte der Versuchspersonen ein, allerdings ist in diesem Fall mit einer Verhaltensänderung zu rechnen, dem so genannten Beobachtungseffekt (vgl. HAMMANN, ERICHSON, 2000, S. 98). Die Messung des Anteils der Personen, der nach einem Toilettenbesuch die Hände wäscht, kann beispielsweise nicht mittels offener Beobachtung erfolgen. Derartige Nachteile versucht die offene aber *nicht durchschaubare* Beobachtung auszugleichen (vgl. MEFFERT, 1992, S. 199). Hier ist den Versuchspersonen zwar bekannt, dass eine Beobachtung vorgenommen wird, das genaue Versuchsziel bleibt aber unbekannt (vgl. MEFFERT, 1992, S. 199).

Bei der Fremdbeobachtung können technische Hilfsmittel, etwa Videokameras, die Rolle des Beobachters übernehmen. Demzufolge unterscheidet man die *persönliche* von der *unpersönlichen* Beobachtung (vgl. BEREKOVEN et al., 1989, S. 118). Die unpersönliche Beobachtung ist von besonderer Bedeutung, wenn bei der Erhebung technische Hilfsmittel zur Aufzeichnung und Speicherung der Daten unumgänglich sind (vgl. MEFFERT, 1992, S. 199). Die Messung des

Anteils der Personen, der nach dem Toilettenbesuch die Hände wäscht, ist nur mit Hilfe von Aufzeichnungsgeräten zu realisieren. Es existieren auch Mischformen, bei denen ein menschlicher Beobachter zusätzlich technische Hilfsmittel einsetzt.

Bei der persönlichen Beobachtung ist die *teilnehmende* von der *nicht-teilnehmenden* Beobachtung zu separieren (vgl. MAYNTZ et al., 1972, S. 90). Bei der nicht-teilnehmenden Beobachtung beschränkt sich der Beobachter lediglich auf die Wahrnehmung der Aktivitäten anderer (vgl. MEFFERT, 1992, S. 199). Die teilnehmende Beobachtung ist demgegenüber durch die aktive Teilnahme des Beobachters charakterisiert (vgl. HAMMANN, ERICHSON, 2000, S. 99). Diese Form der Beobachtung wird eher selten eingesetzt, da die aktive Teilnahme des Beobachters die Objektivität gefährden kann oder eine verdeckte Untersuchungssituation verhindert (vgl. HAMMANN, ERICHSON, 2000, S. 99).

In der Praxis spielt die Beobachtung als Erhebungsmethode eine eher untergeordnete Rolle, da sie lediglich für bestimmte Untersuchungszwecke geeignet ist (vgl. HAMMANN, ERICHSON, 2000, S. 98).

### 3.1.2 Befragung

Da viele Fragestellungen der Wirtschaftswissenschaften nicht offen zu Tage treten und deshalb kaum mittels Beobachtung zu erheben sind, eignet sich in diesem Bereich insbesondere die Befragung als Messinstrument (vgl. HAMMANN, ERICHSSON, 2000, S. 74). Die Befragung ist ein sehr facettenreiches Messinstrument, das ganz allgemein verstanden vom Polizeiverhör bis zur ärztlichen Diagnose reicht (vgl. TOURANGEAU et al., 2005, S. 1).

Im Rahmen einer Befragung sind viele unterschiedliche Fragestellungen vorstellbar. Grundsätzlich können folgende Kategorien unterschieden werden:

- Situationsfragen
- Wertungsfragen

*Situationsfragen* fragen nach konkreten persönlichen Daten oder Aktivitäten der Auskunftsperson, z.B. nach einem bestimmten Verhalten. Bei *Wertungsfragen* stehen dagegen Meinungen, Überzeugungen oder Einstellungen im Mittelpunkt.

Daneben können Befragungen an Hand zweier Gesichtspunkte gegliedert werden:

- Befragungsarten (Kapitel 3.1.2.1)
- Fragetechniken (Kapitel 3.1.2.2)

Bei der Befragungsart sind die Abgrenzung der schriftlichen von der mündlichen Befragung und die Unterscheidung zwischen freier und standardisierter Befragung relevant. Fragetechniken behandeln die konkrete Formulierung der Fragestellung.

### 3.1.2.1 Befragungsarten

Die Durchführung einer Befragung kann sowohl im Rahmen eines persönlichen Interviews, als auch in schriftlicher Form erfolgen (vgl. KALLMANN, 1979, S. 50). Die *schriftliche Befragung* wird mit Hilfe eines Fragebogens ohne direkten Kontakt zwischen Auskunftsperson und Interviewer vorgenommen. Bei der *mündlichen Befragung* stellt ein Interviewer die Fragen und zeichnet die Antworten mit Hilfe von Stichpunkten oder eines Diktiergerätes auf (vgl. HAMMANN, ERICHSON, 2000, S. 78).

Die mündliche Befragung kann in Form eines telefonischen oder persönlichen Interviews erfolgen (vgl. HAMMANN, ERICHSON, 2000, S. 79). Die zusätzliche Unterstützung durch Computer führte in diesem Bereich in den letzten 25 Jahren zu grundlegenden Änderungen (vgl. TOURANGEAU et al., 2005, S. 289). Computergestützte Befragungstechniken, etwa CATI (Computer-Assisted Telephone Interview) oder CAPI (Computer-Assisted Personal Interview) sind eine Verfeinerung persönlicher oder telefonischer Interviews und zählen heute zu den gebräuchlichsten Methoden der Datenerhebung (vgl. TOURANGEAU et al., 2005, S. 289). Der Vorteil dieser Techniken liegt in der Verkürzung der Befragungsdauer, die durch Verzweigungsfragen ermöglicht wird (vgl. GIERL, 1995, S. 209). So muss beispielsweise ein Proband, der angibt kinderlos zu sein, nicht zusätzlich nach dem Alter der Kinder gefragt werden. Davon zu unterscheiden ist die *Computerbefragung*, bei der die Auskunftspersonen die Antworten direkt am Computer eingeben und somit kein Interviewer erforderlich ist (vgl. HAMMANN, ERICHSON, 2000, S. 79). Abbildung 3.2 veranschaulicht die unterschiedlichen Formen der Befragung.

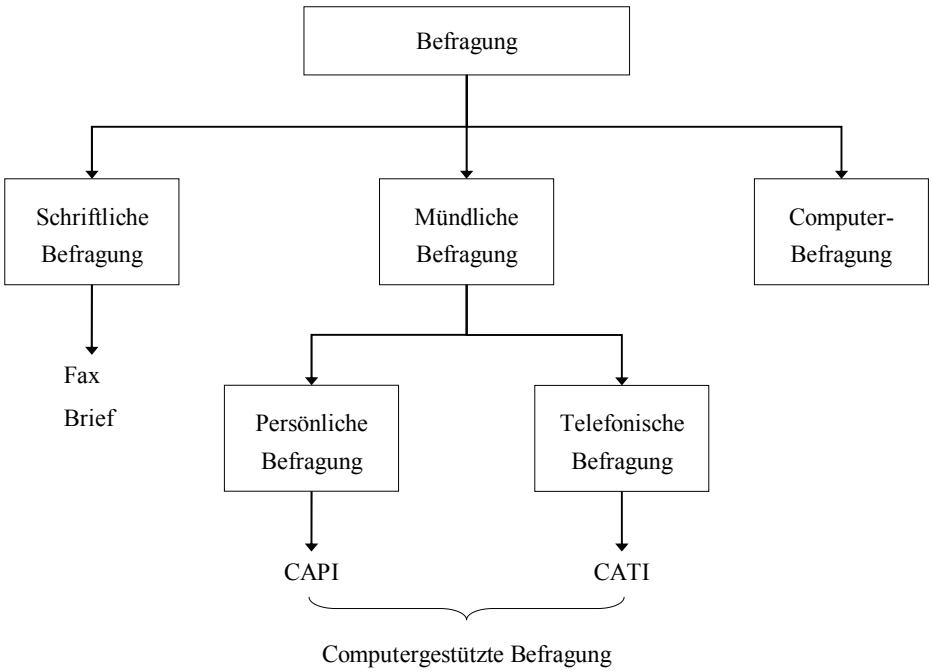


**Abbildung 3.2: Befragungsarten (in Anlehnung an HAMMANN, ERICHSON, 2000, S. 79)**

Eine weitere Unterscheidung kann hinsichtlich der Frageformulierung getroffen werden. Bei einer *standardisierten Befragung* sind die Fragen vorformuliert, während bei der nicht-standardisierten oder *freien Befragung* lediglich ein Leitthema festgelegt ist (vgl. GIERL, 1995, S. 207).

Für die freie Befragung ist ein Interviewer erforderlich, zur Durchführung ist daher lediglich die mündliche Befragung und hier insbesondere die persönliche Befragung geeignet. Da weder die inhaltliche Form noch die Reihenfolge der Fragen vorgegeben ist, steigt das Maß der Verantwortung für den Interviewer, denn dieser trägt entscheidend zur Qualität der erhobenen Messwerte bei (vgl. MEFFERT, 1992, S. 205). Das nicht-standardisierte Interview wird vor allem bei explorativen Fragestellungen angewandt, um für das interessierende Problemfeld Hypothesen zu generieren (vgl. KALLMANN, 1979, S. 50). Da der Einfluss des Interviewers sehr groß ist, eignet sich die nicht-standardisierte Befragung nicht zur Überprüfung bereits formulierter Forschungshypothesen (vgl. GIERL, 1995, S. 209).

Bei der standardisierten Befragung kann sowohl auf die schriftliche, als auch auf die mündliche Befragung zurückgegriffen werden. Bei beiden Erhebungsformen kommen Fragebögen zum Einsatz. Im Rahmen der mündlichen Befragung muss

der Interviewer die Fragen sukzessive abarbeiten und kann somit nur wenig Einfluss auf die Befragung nehmen (vgl. MEFFERT, 1992, S. 205). Durch geeignete Formulierung des Fragebogens dient diese Erhebungstechnik der Überprüfung von Hypothesen. Daher soll an dieser Stelle noch etwas näher auf die Entwicklung von Fragebögen eingegangen werden.

### 3.1.2.2 Fragetechniken

Ein zentrales Thema bei der Erstellung eines Fragebogens ist, ob Antwortkategorien vorgegeben werden oder die Beantwortung offen erfolgen soll (Kapitel 3.1.2.2.1). Ein weiterer Aspekt ist die Art der Frageformulierung, insbesondere die in der Praxis häufig vorkommende Suggestivfrage, welche die Antwort des Probanden beeinflussen soll (Kapitel 3.1.2.2.2).

#### 3.1.2.2.1 Offene vs. geschlossene Fragestellungen

Eine *geschlossene Frage* liegt dann vor, wenn dem Probanden Antwortalternativen vorgegeben sind. Die Auskunftsperson muss damit lediglich die zutreffende Antwortkategorie angeben. Die *offene Frage* verzichtet auf die Angabe von Antwortmöglichkeiten. Die Probanden müssen ihre Antwort selbst formulieren (vgl. MEFFERT, 1992, S. 205).

In vielen Fällen ist die *geschlossene* Frage zur Messung besser geeignet. Durch die Vorgabe der Antwortmöglichkeiten wird sichergestellt, dass die Probanden relevante Antworten für das Untersuchungsziel geben. Die *offene* Fragestellung ist dann geeignet, wenn eine Auflistung der Alternativen die Antwort der Auskunftsperson beeinflusst. Dazu zählt etwa die Prüfung der Gegenwärtigkeit von Wissen (vgl. MAYNTZ et al., 1972, S. 108f.). Unter diesen Aspekt fällt beispielsweise die Frage, wer der aktuelle Außenminister der Bundesrepublik Deutschland ist. Eine geschlossene Fragestellung würde zu einem höheren Bekanntheitsgrad führen als eine offene Fragestellung, weil durch den Ausschluss unmöglicher oder unwahrscheinlicher Alternativen ein Proband selbst dann zur richtigen Antwort gelangen kann, wenn er sich an den Namen des aktuellen Außenministers nicht aktiv erinnern kann.

Das durch geschlossene Fragen gewonnene Skalenniveau ist in der Regel nicht einheitlich. Bei *Auswahlfragen* müssen sich die Probanden für eine von mehreren

verschiedenen Antwortalternativen entscheiden. Die zu Grunde gelegten Antwortkategorien können dabei eine Rangfolge aufweisen oder ungeordnet sein. Weisen die Antwortkategorien eine Rangordnung auf, haben die erhobenen Daten zumindest ordinales Skalenniveau (vgl. HAMMANN, ERICHSON, 2000, S. 86). Unterstellt man zusätzlich Äquidistanz zwischen den Antwortkategorien, so kann sogar von einer Intervallskala ausgegangen werden. Die Äquidistanz der Antwortkategorien muss jedoch kritisch hinterfragt werden, da diese in vielen Fällen nicht nachprüfbar ist.

Sind die Antwortkategorien dagegen ungeordnet, liefert die Auswahlfrage lediglich nominales Skalenniveau. Einen Spezialfall der Auswahlfrage stellt die *Alternativfrage* dar, bei der eine Antwortalternative die andere ausschließt, etwa beim Geschlecht (vgl. HAMMANN, ERICHSON, 2000, S. 85). Dagegen müssen die Probanden bei der *Selektivfrage* eine oder mehrere Kategorien auswählen (vgl. HAMMANN, ERICHSON, 2000, S. 85). Der Vorgabe der Antwortkategorien durch den Untersuchungsleiter kommt eine große Bedeutung zu. So ist die Forderung nahe liegend, dass die zur Verfügung gestellten Kategorien alle möglichen Antworten abdecken müssen (vgl. HOLM, 1975, S. 57). Zur Vermeidung von Ungenauigkeiten sollten weiterhin alle Alternativen relevant sein (vgl. HAMMANN, ERICHSON, 2000, S. 86). Um darüber hinaus Reihenfolgeeffekte auszuschließen, sollte die Abfolge der Kategorien bei den Probanden durchgetauscht werden (vgl. KROSNICK, ALWIN, 1987, S. 205).

#### 3.1.2.2.2 Suggestive Fragen

In der Praxis sind insbesondere bei Meinungsumfragen häufig *Suggestivfragen* vorzufinden. Mit suggestiven Fragen sind Formulierungen gemeint, die den Probanden in seiner Antwort beeinflussen (vgl. MAYNTZ et al., 1972, S. 107f.). Es ist daher offensichtlich, dass auf Suggestivfragen verzichtet werden sollte. Trotzdem sollen sie an dieser Stelle erwähnt werden, da sie häufig als Instrument zur Manipulation des Meinungsbildes eingesetzt werden. Darüber hinaus können Suggestivfragen auch unabsichtlich durch Unachtsamkeit oder ungeschickte Wortwahl entstehen. So zeigen Untersuchungen, dass eine suggestive Frage auch durch die Verwendung von stereotypen Formulierungen entsteht. Bei stereotypen Formulierungen handelt es sich um Wörter oder Wortkombinationen, die in bestimmten Personenkreisen negativ oder positiv besetzt sind (vgl. HOLM, 1975, S. 60). Das positive oder negative Meinungsbild zur Wortkombination überlagert

dann die Meinung zum tatsächlich befragten Thema. Ein Beispiel hiefür könnte derzeit der Begriff ´Globalisierung´ sein.

## 3.2 Messung von Konstrukten

Bei wirtschafts- und sozialwissenschaftlichen Fragestellungen ist die zu messende (interessierende) Variable häufig nicht direkt empirisch erfassbar (vgl. HAMERLE, 1982, S. 7). Möchte man beispielsweise messen, ob eine Person glücklich ist, so könnte dies Ausdruck in einer gefühlten inneren Ruhe der Person finden, dem Ausschütten von Endorphinen im Gehirn, einem Lächeln oder gar einem freudigen Aufspringen (vgl. EID, DIENER, 2006, S. 3). Dennoch ist der Grad des Glücklichseins nicht direkt messbar. Dies ist ein gravierender Unterschied zur Messung im physikalischen Bereich, bei der die interessierenden Merkmale eher direkt beobachtbar sind (vgl. EID, DIENER, 2006, S. 4).

**Definition 3.1**

*Unter einer **latenten Variable (Konstrukt)** sind über einen längeren Zeitraum konstant bleibende Grundcharakteristika von Menschen zu verstehen, etwa Einstellungen oder Fähigkeiten (z.B. Intelligenz), die nicht direkt empirisch messbar sind (vgl. HAMERLE, 1982, S. 7).*

Die Messung von Konstrukten kann nur mit Hilfe beobachtbarer Variablen erfolgen, die mit der latenten Variable in enger Beziehung stehen (vgl. HAMERLE, 1982, S.8).

**Definition 3.2**

*Beobachtbare Merkmale, die zur Messung von latenten Variablen herangezogen werden, heißen **manifeste Variable (Indikatoren).** Werden die manifesten Variablen mit Hilfe einer direkten Befragung erhoben, so spricht man von **Items**. Alternativ können zur Messung auch **Objekte (Stimuli)** mit bestimmten Eigenschaften herangezogen werden, die von den Personen bewertet werden. Insbesondere im Bereich der Psychologie kann schließlich auch die Fähigkeit zur **Lösung bestimmter Aufgaben (Tests)** als Indikator verwendet werden (vgl. HAMERLE, 1982, S. 8 und MEFFERT, 1992, S. 57).*

Diese Definition macht deutlich, dass eine weitere Präzisierung der in Kapitel 2 definierten Begriffe ´Objekt´ und ´Merkmal´ erforderlich ist. Insbesondere wurde in Kapitel 2 keine Unterscheidung von Objekten und Auskunftspersonen vorgenommen. Häufig entspricht bei der wirtschafts- und sozialwissenschaftlichen Messung die Auskunftsperson jedoch nicht dem Beurteilungsobjekt. Die Identität von Objekt und Auskunftsperson wäre beispielsweise bei der Frage nach der Körpergröße der Auskunftsperson gegeben. Soll der Proband dagegen aus fünf Automarken auswählen, welche er präferiert, sind Auskunftsperson und Beurteilungsobjekt verschieden.

Ein Beispiel für eine Konstruktmessung mit Hilfe von beobachtbaren Variablen sind Intelligenztests. Die latente Variable Intelligenz ist nicht direkt beobachtbar. Als Indikatoren zur Messung der Intelligenz dienen verschiedene Tests, welche dann zu einer Gesamtpunktzahl addiert werden. Die Messung von Konstrukten ist jedoch häufig mit großen Schwierigkeiten verbunden, wie folgendes Beispiel von NELSON, GRINDER und MUTTERER (1969) verdeutlicht.

**Beispiel 3.1**

*In ihrer Untersuchung befassen sich die Autoren mit der Messung des Konstrukts ´Ehrlichkeit´. Als Indikatoren verwenden die Verfasser sechs Testsituationen. Jede dieser Testsituationen soll das Konstrukt ´Ehrlichkeit´ messen. Details der Tests ´Ray Gun´, ´Magic Mirror´, ´Multiple Choice´, ´Speed´, ´Squares´ und ´Circles´ sind für das grundlegende Verständnis der Problematik nicht erforderlich.*[1] *In der Studie werden insgesamt 106 Schüler überprüft, die jeweils an allen Tests teilnehmen. Die Schüler entsprechen somit den Auskunftspersonen der Messung. Das wahre Untersuchungsziel, die Messung der Ehrlichkeit, ist den Kindern nicht bekannt. Den Jugendlichen wird suggeriert, dass mit Hilfe der sechs Tests ihre Fähigkeit in bestimmten Bereichen, etwa das räumliche Vorstellungsvermögen bei ´Squares´ und ´Circles´, geprüft wird. Die erreichten Punktzahlen bei den einzelnen Tests entsprechen diesen Fähigkeiten. Die Ehrlichkeit der Schüler wird folgendermaßen überprüft: Die erreichten Punktzahlen bei den Tests dürfen die Schüler selbst mitteilen, d.h., die Schüler dürfen ihre eigene Arbeit an Hand einer Musterlösung korrigieren. Anschließend prüft der Untersuchungsleiter, ob die Schüler wahrheitsgemäße Ergebnisse berichten. Dies wissen die Schüler jedoch nicht.*

[1] Die einzelnen Tests sind bei NELSON, GRINDER und MUTTERER (1969) ausführlich beschrieben.

*Zusätzlich schaffen die Autoren bei den einzelnen Tests unterschiedliche Anreize für die Jugendlichen, erhöhte Punktzahlen und somit unehrliche Ergebnisse zu berichten.*

*Falls die sechs Testsituationen gleichermaßen einen Indikator für das Konstrukt ´Ehrlichkeit´ darstellen, sollten alle Tests zu vergleichbaren Ergebnissen führen. Bei jedem Test kann eine 0/1-codierte Variable gebildet werden, die angibt, ob der Schüler eine überhöhte Punktzahl angegeben hat oder nicht. Für die auf diese Weise bei jedem Test gebildeten Variablen kann die Korrelation berechnet werden. Diese Korrelationen müssten sehr hoch sein, wenn alle Tests geeignete Indikatoren für das Konstrukt ´Ehrlichkeit´ sind. Dabei ist zu beachten, dass auf Grund der unterschiedlichen Motivation die relative Häufigkeit der betrügenden Schüler zwischen den Tests variieren kann (vgl. NELSON et al., 1969, S. 269). Tabelle 3.1 zeigt die Anzahl der Schüler, die bei den einzelnen Tests ehrlich bzw. unehrlich sind:*

| | *Ray Gun* | *Magic Mirror* | *Multiple Choice* | *Speed* | *Squares* | *Circles* |
|---|---|---|---|---|---|---|
| *Ehrlich* | *15* | *46* | *39* | *81* | *41* | *58* |
| *Unehrlich* | *91* | *60* | *67* | *25* | *65* | *48* |

**Tabelle 3.1: Anzahl der ehrlichen/unehrlichen Schüler (vgl. NELSON et al., 1969, S. 270)**

*Wie zu erwarten, bestehen bei den Tests Unterschiede bezogen auf die Anzahl der unehrlichen Schüler. Allerdings erwartet man von einem sehr ´ehrlichen´ Schüler, dass er bei allen Tests wahrheitsgemäß berichtet. Ein ´unehrlicher´ Schüler wird tendenziell bei allen Tests eine falsche Punktzahl angeben. Bei Kindern zwischen diesen beiden Extremgruppen sollte davon ausgegangen werden, dass sie in jenen Testsituationen die Wahrheit berichten, die keinen starken Anreiz zum Lügen geben. In Testsituationen mit hoher Motivation zur Angabe überhöhter Punkte ist dagegen ein unehrliches Verhalten zu erwarten.*

*Jene 15 Kinder, die bei ´Ray Gun´ ehrliche Angaben machen, sollten bei allen anderen Tests ebenfalls ehrlich sein. Dagegen sollten die 25 Schüler, die bei ´Speed´ unehrlich sind bei sämtlichen anderen Tests ebenfalls lügen. Da beim Vergleich von ´Ray Gun´ und ´Speed´ maximal 40 Schüler übereinstimmende Testergebnisse vorweisen, kann für diese beiden Tests keine sehr hohe Korrelation entstehen. Liefern zwei Tests, beispielsweise ´Multiple Choice´ und*

*´Squares´ jedoch sehr ähnliche Ergebnisse hinsichtlich der Häufigkeiten, dann sollte die Korrelation sehr hoch sein. Bei der Berechnung der Korrelationen für die Tests erhält man die Ergebnisse in Tabelle 3.2:*

| | *Ray Gun* | *Magic Mirror* | *Multiple Choice* | *Speed* | *Squares* | *Circles* |
|---|---|---|---|---|---|---|
| *Ray Gun* | - | *0,21* | *0,40* | *0,18* | *0,07* | ***-0,08*** |
| *Magic Mirror* | *0,21* | - | *0,33* | *0,13* | *0,43* | *0,28* |
| *Multiple Choice* | *0,40* | *0,33* | - | *0,59* | *0,34* | *0,42* |
| *Speed* | *0,18* | *0,13* | *0,59* | - | ***-0,10*** | *0,01* |
| *Squares* | *0,07* | *0,43* | *0,34* | *-0,10* | - | *0,77* |
| *Circles* | *-0,08* | *0,28* | *0,42* | *0,01* | *0,77* | - |

**Tabelle 3.2: Korrelation zwischen den einzelnen Tests (vgl. NELSON et al., 1969, S. 272)**

*Die teilweise sogar negativen Korrelationen verdeutlichen, dass die Indikatoren offensichtlich nicht alle übereinstimmend das Konstrukt ´Ehrlichkeit´ messen bzw. dass nicht alle Indikatoren vom Konstrukt ´Ehrlichkeit´ gleichermaßen beeinflusst werden. Die Tests messen somit unterschiedliche Aspekte ehrlichen Verhaltens.*

Das Beispiel des Konstrukts ´Ehrlichkeit´ verdeutlicht, wie komplex die Messung latenter Variablen sein kann. Die Ergebnisse der Studie legen den Schluss nahe, dass die Ehrlichkeit einer Person situationsbezogen ist. Demzufolge müssen verschiedene Gesichtspunkte bei der Beurteilung der Ehrlichkeit einer Person berücksichtigt werden.

Allgemein spricht man von der Mehrdimensionalität eines Konstrukts. Die verschiedenen Dimensionen eines Konstrukts können mit Hilfe von Items gemessen werden, wie Abbildung 3.3 veranschaulicht. Beim Konstrukt Intelligenz entsprechen die verschiedenen Dimensionen zum Beispiel der mathematischen, sprachlichen und künstlerischen Begabung einer Person. Diese Dimensionen werden mit Hilfe von Indikatoren gemessen. Bei der Intelligenz-messung erfolgt dies durch Tests, bei anderen Fragestellungen können auch Items oder Objektbewertungen eingesetzt werden. Zur Messung einer Dimension eines Konstrukts können auch mehrere Indikatoren herangezogen werden, wie bei Dimension 3 in Abbildung 3.3 angedeutet.

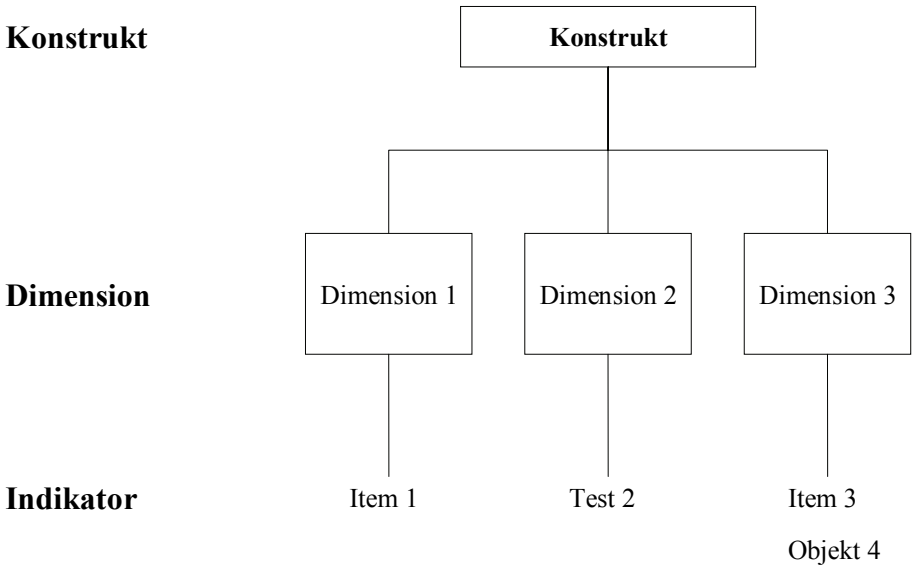


**Abbildung 3.3: Mehrdimensionalität eines Konstrukts**

In den Wirtschaftswissenschaften ist insbesondere die Einstellungsmessung ein zentrales Forschungsgebiet, das auf die Messung von Konstrukten zurückzuführen ist (vgl. GIERL, 1995, S. 33-210). Eine Einstellung entsteht durch Lernprozesse, etwa durch direkte oder indirekte Erfahrungen mit einem Objekt (vgl. MEFFERT, 1992, S. 55). Jede Einstellung kann aus einer oder mehreren Dimensionen bestehen. Bei der Messung einer Dimension der Einstellung, sind drei Komponenten zu berücksichtigen (vgl. HAMMANN, ERICHSON, 2000, S. 267):

- Affektive (gefühlsmäßige) Komponente
- Kognitive (wissensbasierte) Komponente
- Konative (intentionale) Komponente

Manche Autoren sehen die konative Komponente (Intention) als Folge der Einstellung und somit nicht als ihren Bestandteil an (vgl. KROEBER-RIEL, WEINBERG, 1999, S. 167). Ob eine Versuchsperson eher eine positive oder negative Einstellung zum Leben hat, hängt sowohl von affektiven als auch kognitiven Komponenten ab. Das Wissen, in Westeuropa ein Leben in Luxus führen zu können, stellt eine kognitive Komponente dar. Die Trauer um einen vor kurzem verstorbenen Angehörigen dagegen eine affektive Komponente.

Auf Grund der Mehrdimensionalität der Einstellung und weil bei der Messung der Dimensionen affektive und kognitive Komponenten zu berücksichtigen sind, kann eine Einstellung nur über Umwege gemessen werden. Dies soll für ein eindimensionales Konstrukt veranschaulicht werden (vgl. Abbildung 3.4):

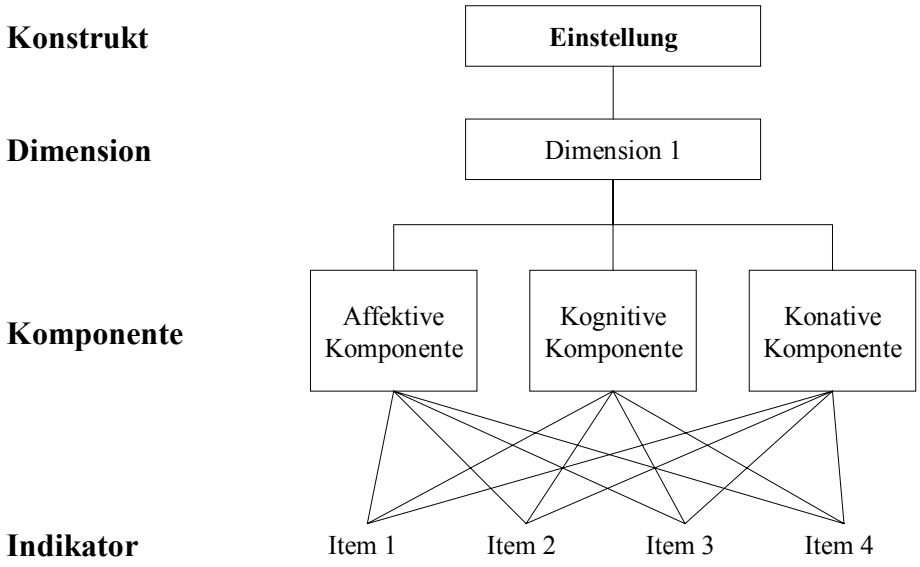


**Abbildung 3.4: Messung des Konstrukts Einstellung (in Anlehnung an KROEBER-RIEL, WEINBERG, 1999, S. 184)**

Diese einleitenden Erklärungen verdeutlichen, weshalb Messungen im wirtschafts- und sozialwissenschaftlichen Bereich wesentlich komplexer als im physikalischen Bereich sind. Die Wahl der Erhebungsmethode in Kapitel 3.1 bezieht sich auf die Entscheidung, mit welchen Mitteln die Indikatoren zu erheben sind. Die in Kapitel 3.3 behandelten Skalierungsverfahren versuchen die Beziehung zwischen Konstrukt, Dimension und Indikator geeignet durch Messvorschriften abzubilden.

## 3.3 Skalierungsverfahren

Skalierungsverfahren behandeln die konkrete Operationalisierung der Messung, die insbesondere bei Konstrukten auf erhebliche Schwierigkeiten stößt. Da die Messung von Konstrukten in der Regel mittels einer Befragung erfolgt, sollen hier lediglich Skalierungsverfahren für Befragungen vorgestellt werden (vgl.

KALLMANN, 1979, S. 57). Je nachdem, wie viele Dimensionen die zu Grunde gelegte latente Variable aufweist, unterscheidet man eindimensionale und mehrdimensionale Verfahren (vgl. KALLMANN, 1979, S. 59). Beide Verfahren greifen dabei auf dieselben grundlegenden Fragetechniken zurück.

- Die *grundlegenden Verfahren* sind Ratingskalen und die Methode der Paarvergleiche. Auf einem dieser beiden Typen bauen im Prinzip alle Skalierungsverfahren auf. Daher sollen in Kapitel 3.3.1 zunächst diese beiden Verfahren vorgestellt werden.

- *Eindimensionale Skalierungsverfahren* weisen dem mit Ratingskalen oder Paarvergleichen gemessenen Sachverhalt genau einen Wert zu. Sie sind somit nicht in der Lage, die Mehrdimensionalität von Einstellungen geeignet abzubilden. Dennoch liefern sie zum Teil brauchbare Ergebnisse in der Konstruktmessung. Sie werden in Kapitel 3.3.2 näher beschrieben.

- *Mehrdimensionale Skalierungsverfahren* versuchen verschiedene Dimensionen eines Konstrukts zu messen (Kapitel 3.3.3). Ein Teil der Verfahren, etwa die Modelle von FISHBEIN und TROMMSDORFF, berechnet für mehrdimensionale Konstrukte einen einzigen konkreten Wert. Die Mehrdimensionalität wird lediglich durch Aggregationsregeln zum Ausdruck gebracht. Ein anderer Teil bildet die Mehrdimensionalität des Konstrukts durch Zuweisung eines Zahlentupels ab. Beide Verfahrenstypen basieren wiederum auf Ratingskalen oder Paarvergleichen.

Abbildung 3.5 veranschaulicht die Einteilung der Skalierungsverfahren:

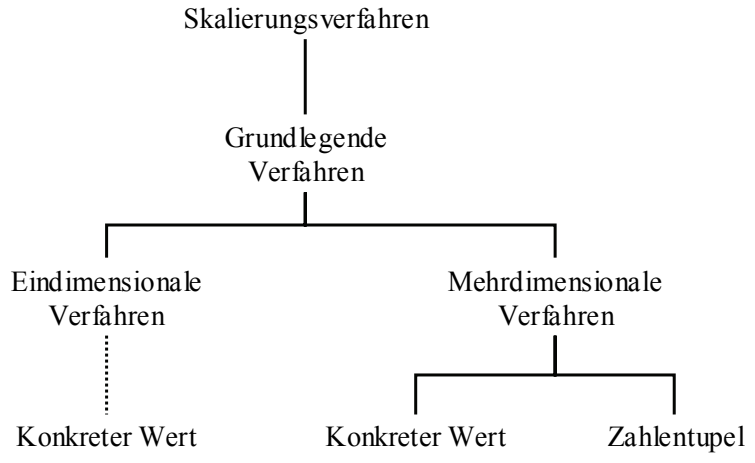


**Abbildung 3.5: Einteilung der Skalierungsverfahren**

## 3.3.1 Grundlegende Verfahren

Zur Messung der beobachtbaren Variablen müssen Probanden meist den Grad der Zustimmung bei bestimmten Items angeben (vgl. HAMMANN, ERICHSON, 2000, S. 273). Die Personen können ihren individuellen Grad der Zustimmung entweder mittels Ratingskalen direkt angeben oder indirekt durch die Angabe von Präferenzen bei Paarvergleichen (vgl. GIERL, 1995, S. 53f.). Diese beiden Methoden bilden die Grundlage, auf der alle anderen eindimensionalen Skalierungsverfahren aufgebaut sind, die in den weiteren Abschnitten vorgestellt werden.

### 3.3.1.1 Ratingskalen

Die am häufigsten eingesetzte Methode zur Messung von Einstellungen ist die *Ratingskala* (vgl. HAMMANN, ERICHSON, 2000, S. 274). Bei ihr handelt es sich um eine Form der geschlossenen Frage. Die Antwortkategorien sind entweder mit Zahlen oder verbal beschriftet (vgl. GIERL, 1995, S. 44). In Abbildung 3.6 ist ein Beispiel für eine Ratingskala aufgeführt:

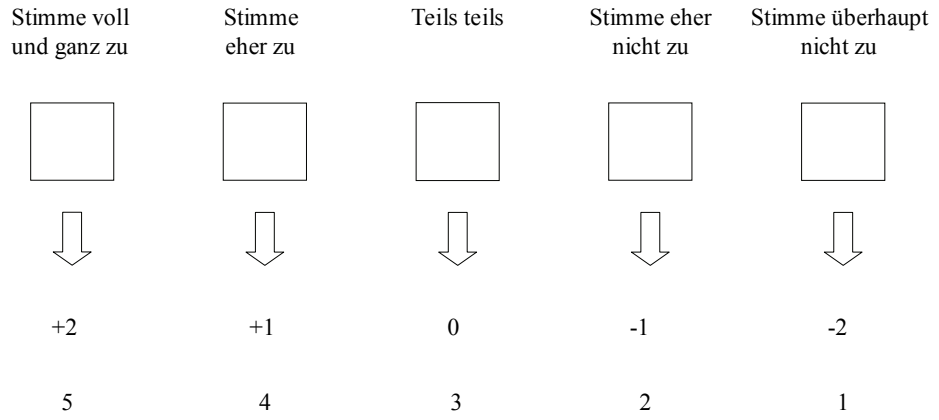


**Abbildung 3.6: Beispiel für eine Ratingskala**

Den jeweiligen Kategorien wird eine geeignete Zahl zugeordnet, welche dem Skalenwert des Probanden bei diesem Item entspricht. Da bei der Zuordnung der Zahl lediglich die Rangfolge der Ausprägungen wesentlich ist, messen

Ratingskalen prinzipiell auf ordinalem Skalenniveau (vgl. MEFFERT, 1992, S. 185). Unterstellt man jedoch Äquidistanz zwischen den Kategorien, so erfüllen Ratingskalen alle Voraussetzungen einer Intervallskala (siehe Seite 15). Inwieweit die Annahme der Äquidistanz empirisch belegbar ist, liegt in den meisten Fällen im „Ermessen des Untersuchenden“ (HAMMANN, ERICHSON, 2000, S. 86). Da bei intervallskalierten Merkmalen deutlich mehr numerische Aussagen möglich sind als bei ordninalskalierten, sehen viele Autoren die Ratingskala als intervallskaliert an. Dies ist jedoch nicht empirisch begründet, sondern basiert auf subjektiven Annahmen bzw. den bekannten Vorteilen von Intervallskalen.

Bei einer Ratingskala müssen verschiedene Gesichtspunkte berücksichtigt werden, insbesondere die beiden folgenden:

- Anzahl der Antwortkategorien
- Beschriftung der Antwortkategorien

Die *Anzahl der Antwortkategorien* ist ein zentraler Aspekt der Ratingskala. Da dieser Bereich ausführlich in Kapitel 5 und 6 dieser Arbeit beschrieben wird, soll an dieser Stelle nicht näher darauf eingegangen werden.

Völlig unabhängig von der Zahl der Antwortkategorien ist deren *Beschriftung*. Dadurch wird unter anderem die Polarität der Skala festgelegt. Bei einer *bipolaren* Skala sind die Skalenenden mit gegenteiligen Ausprägungen beschriftet, etwa ´ist sehr groß´ und ´ist sehr klein´ (vgl. GIERL, 1995, S. 45). Die *unipolare* Skala ist durch unterschiedliche Zustimmungsgrade, beispielsweise von ´stimme überhaupt nicht zu´ bis ´stimme voll und ganz zu´, gekennzeichnet (vgl. GIERL, 1995, S. 45). Die Beschriftung einer Skala ist vor allem deshalb bedeutsam, weil durch sie die Äquidistanz der Messwerte sichergestellt werden soll. Diese Äquidistanz erlaubt es nämlich, die so gewonnenen Daten als intervallskaliert zu behandeln. Eher technischer Natur ist die Frage, ob lediglich die Endpunkte der Skala beschriftet werden sollen, oder ob alle Antwortkategorien verbal beschrieben werden.

In der Literatur sind einige Phänomene diskutiert, die vom Probanden nicht beabsichtigte Antwortverzerrungen bei Ratingskalen zur Folge haben (vgl. HAMMANN, ERICHSON, 2000, S. 274f.). Dazu zählt der *Nachsichtseffekt*, der beschreibt, dass bekannte Untersuchungsobjekte besser beurteilt werden als

unbekannte. Der *Zentralitätseffekt*, wonach Personen extreme Beurteilungen vermeiden und der *Haloeffekt*, dass Probanden bei der Bewertung von Objekten durch übergeordnete Sachverhalte beeinflusst werden. So überlagert die Einstellung zum Land unter Umständen die Einstellung zu italienischen Weinen.

### 3.3.1.2 Paarvergleiche

Neben den Ratingskalen existieren auch Skalierungsmethoden, die auf *Paarvergleichen* beruhen. Bei Paarvergleichen werden den Versuchspersonen Objekte (Stimuli) vorgelegt. Für diese Objekte geben die Personen dann eine Präferenz der Form ´Stimulus A ist besser als Stimulus B´ ($A \succ B$) an (vgl. hierzu und zum folgenden Absatz: KALLMANN, 1979, S. 64ff.). Alternativ können die Probanden auch zwei Objektpaare beurteilen. Sie müssen dann entscheiden, welches Stimulipaar sich ähnlicher ist. Während bei der Ratingskala einzelne Items direkt beurteilt werden, bewerten die Probanden bei dieser Methode die Objekte im Vergleich zueinander.

Dies ist für die Auskunftsperson häufig leichter zu bewerkstelligen, da sie ohne weiteres angeben kann, welches Produkt sie präferiert. Beim Vergleich verschiedener Biermarken kann der Proband angeben, welche Biermarke er bevorzugt. Welche Produkteigenschaften, z.B. Alkoholgehalt und Kohlensäuregehalt, dafür verantwortlich sind, kann er dagegen nicht angeben. Allerdings weist die Paarvergleichsmethode auch Nachteile auf. Zum einen steigt der Erhebungs- und Auswertungsaufwand deutlich. Bei n zu bewertenden Objekten müssen die Probanden n(n-1)/2 Paarvergleiche vornehmen, für sieben Stimuli entspricht dies bereits 21 Vergleichen. Außerdem ist für den Untersuchungsleiter unter Umständen nicht nachvollziehbar, weshalb der Proband eine bestimmte Objektrangfolge wählt. In der empirischen Praxis zeigt sich vor allem noch ein dritter Nachteil, der die Zuweisung eines Skalenwertes zu den Objekten deutlich erschwert.

Die Zuweisung von Skalenwerten ist unproblematisch, wenn die Objekte eine vollständige Präordnung ergeben. Dieser Fall liegt vor, wenn bei der Bewertung durch den Probanden keine Intransitivität der Form $(A \succ B \wedge B \succ C \wedge C \succ A)$ (vgl. OPITZ, 2004, S. 121) entsteht. Falls Intransitivitäten auftreten, ist die Zuweisung eines Skalenwertes nicht mehr eindeutig. In diesen Fällen kann behelfsmäßig die Häufigkeit mit der ein Objekt präferiert wird als Skalenwert

verwendet werden (vgl. KALLMANN, 1979, S. 66). Eine Reduktion des Erhebungsaufwands ist möglich, wenn Personen direkt eine Rangfolge der Stimuli oder Objektpaare mitteilen (vgl. KALLMANN, 1979, S. 66). In diesem Fall ist die Paarvergleichsmethode der Ratingskala ähnlich (vgl. KALLMANN, 1979, S. 66). Ein Unterschied ist jedoch, dass bei der Ratingskala zwei Objekte identisch bewertet werden können, was bei der Bildung einer echten Rangfolge nicht möglich ist. Welche der beiden Methoden besser ist, kann nicht allgemein beurteilt werden. Der Zwang für die Probanden, eine Rangfolge auch für solche Objekte anzugeben, die sie nicht unterscheiden können, kann zu schlechteren Ergebnissen der Paarvergleichsmethode führen. Andererseits verleitet die Ratingskala Probanden unter Umständen dazu, selbst solche Objekte identisch zu bewerten, die sie bei näherem Nachdenken in eine Rangfolge hätten bringen können.

### 3.3.2 Eindimensionale Skalierungsverfahren

Bei eindimensionalen Skalierungsverfahren wird nur eine Dimension des interessierenden Konstrukts gemessen. Dabei können durchaus mehrere verschiedene Indikatoren zum Einsatz kommen. Die eindimensionalen Verfahren zeichnen sich jedoch dadurch aus, dass sie diese unterschiedlichen Indikatoren zu einer Kennzahl aggregieren. Daher eignen sie sich nicht für die Messung mehrdimensionaler Konstrukte. Folgende Verfahren basieren auf Ratingskalen:

- Likert-Skala (Kapitel 3.3.2.1)
- Guttman-Skala (Kapitel 3.3.2.2)

Neben diesen beiden Verfahren existieren eindimensionale Verfahren, die auf die Methode der Paarvergleiche zurückgreifen:

- Law of Comparative Judgement (Kapitel 3.3.2.3)
- Unfolding-Technik (Kapitel 3.3.2.4)

### 3.3.2.1 Likert-Skala

Die *Likert-Skala* basiert auf einer Ratingskala, welche als intervallskaliert aufgefasst wird (vgl. HAMMANN, ERICHSON, 2000, S. 275). Jede Auskunftsperson muss dabei für verschiedene Aussagen zu einem Untersuchungsgegenstand Stellung beziehen, die Likert-Skala ist daher eine Multi-Item-Messung (vgl. GIERL, 1995, S. 54). Zunächst erfolgt a priori eine Aufteilung der Items in zwei gleich große Gruppen, die beide das gleiche Thema erfragen. Eine der beiden Gruppen umfasst ausschließlich positiv formulierte Aussagen, während negativ formulierte Items die andere Gruppe bilden (vgl. HAMMANN, ERICHSON, 2000, S. 276). Die Versuchspersonen müssen für jedes Item mit Hilfe einer Ratingskala den Grad ihrer Zustimmung angeben. Somit ist eine Kontrolle der Ergebnisse möglich, da positiv und negativ formulierte Items zu ähnlichen Ergebnissen führen sollten. Die Items werden anschließend aggregiert, wobei bei der Zuweisung von Skalenwerten zu den einzelnen Antwortkategorien der Items folgendes berücksichtigt werden muss:

- Veränderung des Vorzeichens bei negativ formulierten Items
- Äquidistanz der Skalenwerte, um die Aggregation zu rechtfertigen

Das Ergebnis dieser ungewichteten Aggregation ist ein vorläufiger Gesamtpunktwert für jede Person. Auf Basis aller ermittelten Gesamtpunktwerte werden zwei Extremgruppen der Personen gebildet, meist durch das erste bzw. vierte Quartil der Verteilung der Gesamtpunktwerte (vgl. DENZ, 1976, S. 96). Für diese Extremgruppen können mit Hilfe der Korrelationsanalyse und durch Signifikanztests alle trennscharfen Items bestimmt werden (vgl. DENZ, 1976, S. 97). Items, die keine Trennschärfe besitzen, werden bei der weiteren Analyse nicht berücksichtigt. Dieser Ausschluss von Items hat zur Folge, dass die Likert-Skala nur für eindimensionale Konstrukte sinnvolle Ergebnisse liefert, da Items, die weitere Dimensionen des interessierenden Konstrukts betreffen, nicht in das Endergebnis einfließen (vgl. DENZ, 1976, S. 100).

Der endgültige Einstellungswert berechnet sich durch eine ungewichtete Aggregation der trennscharfen Items oder durch eine erneute Befragung ohne die nicht-trennscharfen Items (vgl. HAMMANN, ERICHSON, 2000, S. 276):

$$E_i = \sum_{j=1}^{n} x_{ij} \tag{3.1}$$

$E_i$: *Einstellungswert von Person i*

$x_{ij}$: *Rating von Person i bezüglich Item j*

### 3.3.2.2 Guttman-Skala

Bei der *Guttman-Skala* erfolgt nicht nur eine Skalierung der Items, sondern gleichzeitig auch eine Skalierung der Personen (vgl. HAMMANN, ERICHSON, 2000, S. 278). Diese Skala ist ursprünglich dem Bereich der Messung von Fähigkeiten zuzuordnen, die grundlegenden Charakteristika sind aber auf die Einstellungsmessung übertragbar (vgl. TORGERSON, 1958, S. 307). In Analogie zur Likert-Skala werden den Auskunftspersonen verschiedene Aussagen zu einem eindimensionalen Konstrukt vorgelegt (vgl. SIXTL, 1976, S. 42). Allerdings müssen die Probanden nicht den Grad der Zustimmung, sondern lediglich ihre generelle Zu- oder Ablehnung einer Aussage angeben (TORGERSON, 1958, S. 309). Im Rahmen der Messung von Fähigkeiten der Person entsprechen die Items einzelnen Testsituationen (vgl. SIXTL, 1976, S. 42). Bei der Einstellungsmessung sind die Items wie bei der Likert-Skala in zwei Gruppen eingeteilt, eine Itemgruppe mit positiv formulierten Fragen und eine mit negativ formulierten Fragen. Dabei ist Zustimmung bei positiv formulierten Aussagen mit 1, bei negativ formulierten Items mit 0 codiert (vgl. HAMMANN, ERICHSON, 2000, S. 279).

$$E_i = \sum_{j=1}^{n} x_{ij} \tag{3.2}$$

$E_i$: *Einstellungswert von Person i*

$x_{ij}$: *0/1-codierte Aussage von Person i bezüglich Item j*

Die Addition liefert einen eindimensionalen Einstellungswert, dessen Berechnung in einem Skalogramm dargestellt werden kann (vgl. TORGERSON, 1958, S. 308). Tabelle 3.3 verdeutlicht eine weitere typische Eigenschaft der Guttman-Skala. Man geht bei der Guttman-Skala davon aus, dass eine simultane Rangordnung der Personen und Items überschneidungsfrei möglich ist (vgl. SIXTL, 1976, S. 44).

In Tabelle 3.3 sind vier Items und die Zustimmung (1) oder Ablehnung (0) dieser Items durch 8 Versuchspersonen angegeben. Die Items sind so angeordnet, dass zunächst das Item abgetragen ist, bei dem die wenigsten Personen zugestimmt haben (Item 1). Es folgt das Item mit der zweitgeringsten Zustimmung (Item 2) und anschließend die Items 3 und 4. Entsprechend sind auch die Personen je nach der Anzahl ihrer Zustimmungen geordnet. Die Guttman-Skala fordert, dass eine Person dem weniger radikal formulierten Item 4 (leichterer Test) immer zustimmt, wenn sie auch die radikalere Aussage in Item 3 (komplizierterer Test) bejaht (vgl. TORGERSON, 1958, S. 310). Daher ist aus der Anzahl der bejahten Items (gelösten Aufgaben) eindeutig ersichtlich, welchen Aussagen (Tests) der Proband zugestimmt hat.

| Person | Items | | | | | | | | Einstellungswert |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | | (2) | | (3) | | (4) | | |
| | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | |
| 1 | x | | x | | x | | x | | 4 |
| 2 | | x | x | | x | | x | | 3 |
| 3 | | x | x | | x | | x | | 3 |
| 4 | | x | | x | x | | x | | 2 |
| 5 | | x | | x | x | | x | | 2 |
| 6 | | x | | x | | x | x | | 1 |
| 7 | | x | | x | | x | x | | 1 |
| 8 | | x | | x | | x | | x | 0 |

**Tabelle 3.3: Beispiel für ein Skalogramm (in Anlehnung an HAMMANN, ERICHSON, 2000, S. 279)**

Selbst im Rahmen der Fähigkeitenmessung dürfte diese Eigenschaft der Guttman-Skala nur in Ausnahmefällen empirisch haltbar sein. Wie schwierig ein Test für eine Person ist, hängt entscheidend von den individuellen Fähigkeiten ab. Vergleichbares lässt sich auch für die Beurteilung der Items bei der Einstellungsmessung vermuten. Die idealtypische Einteilung der Personen und Items nach Tabelle 3.3 ist in der Praxis selten möglich. Um trotz dieses Kritikpunktes die Verwendung der Guttman-Skala zu rechtfertigen, wird bei der Einstellungsmessung von der Annahme der überschneidungsfreien Anordnung der Personen und Items abgesehen. Stattdessen wird lediglich gefordert, dass Personen und Items so anzuordnen sind, dass sie das theoretische Ideal in Tabelle 3.3 bestmöglich wiedergeben (vgl. TORGERSON, 1958, S. 319).

Zur Messung mehrdimensionaler Konstrukte ist die Guttman-Skala ungeeignet, da dann eine objektive Rangordnung der Items (Tests) und Personen unmöglich ist. Unterstellt man, dass die Intelligenz aus einer mathematischen und einer sprachlichen Komponente besteht, wird diese Problematik deutlich. Es ist davon auszugehen, dass Personen die Schwierigkeit der Multiplikation zweier Zahlen einerseits und des Konjugierens eines Verbs andererseits unterschiedlich wahrnehmen. Innerhalb einer Dimension ist die objektive Rangordnung der Items (Tests) dagegen tendenziell möglich. So liegt zum Beispiel die Vermutung nahe, dass die Multiplikation zweier Zahlen für alle Personen schwieriger ist als deren Addition.

#### 3.3.2.3 THURSTONE´S Law of Comparative Judgement

Die Basis für das *Law of Comparative Judgement* bilden Paarvergleiche von Beurteilungsobjekten (vgl. THURSTONE, 1927, S. 273). Ähnlich wie bei den Items der Guttman-Skala wird davon ausgegangen, dass jedes Beurteilungsobjekt einheitlich für alle Personen eine bestimmte Position auf einem Kontinuum einnimmt (vgl. KALLMANN, 1979, S. 67). Im Gegensatz zur Guttman-Skala lässt das Law of Comparative Judgement jedoch ausdrücklich das Auftreten von inhomogenen Bewertungen durch eine Person zu, z.B. hervorgerufen durch Aufmerksamkeitsschwankungen (vgl. HAMERLE, 1982, S. 20). Die einzelnen Bewertungen eines Stimulus i bilden eine normalverteilte Zufallsvariable $X_i$ mit dem Erwartungswert $\mu_i$ und der Standardabweichung $\sigma_i$ (vgl. THURSTONE, 1927, S. 278). Eine Person entscheidet sich für die Bewertung $i \succ j$ (Objekt i ist besser als Objekt j) wenn $X_i - X_j > 0$ ist (vgl. GIERL, 1995, S. 609). Auf Grund der Annahme der Normalverteilung für $X_i$ und $X_j$ gilt für die Zufallsvariable $X_i - X_j$ Folgendes (vgl. GIERL, 1995, S. 609):

$$X_i - X_j \sim N(\mu_i - \mu_j,\ \sigma_i^2 + \sigma_j^2 - 2\rho_{ij}\sigma_i\sigma_j) \qquad (3.3)$$

$\rho_{ij}$ : *Korrelation zwischen* $X_i$ *und* $X_j$

Allerdings sind alle in (3.3) verwendeten Größen in der Realität unbekannt. Es kann lediglich empirisch beobachtet werden, wie viele Personen das Urteil $i \succ j$ abgeben. Der Anteil der Personen mit diesem Urteil sei $p_{ij}$. Daraus lässt sich das von THURSTONE formulierte Law of Comparative Judgement ableiten (1927, S. 276):

$$\mu_i - \mu_j = z_{ij}\sqrt{\sigma_i^2 + \sigma_j^2 - 2\rho_{ij}\sigma_i\sigma_j} \qquad (3.4)$$

$z_{ij}$: *Standardisierte Werte für* $p_{ij}$

$p_{ij}$: *Empirisch beobachtbarer Anteil der Personen mit dem Urteil* $(i \succ j)$

Löst man (3.4) nach $z_{ij}$ auf, erkennt man, dass der empirisch beobachtbare Anteil der Personen mit dem Urteil i ≻ j von drei Größen abhängt. Unterstellt man, dass $\mu_i$ - $\mu_j$ positiv ist, können folgende Aussagen über den Einfluss dieser Größen getroffen werden:

- $z_{ij}$ steigt, wenn die Distanz der Erwartungswerte ($\mu_i$ und $\mu_j$) zunimmt. Anschaulich bedeutet dies, dass die Bewertung der Objekte i und j homogener wird, wenn sie auf dem psychologischen Kontinuum weiter voneinander entfernt sind (vgl. KALLMANN, 1979, S. 68).

- $z_{ij}$ steigt, wenn die Varianzen $\sigma_i^2$ und $\sigma_j^2$ abnehmen. Demzufolge wird die Bewertung der Objekte i und j homogener, wenn die Einzelbewertungen der Objekte weniger streuen.

- $z_{ij}$ steigt, wenn die Korrelation $\rho_{ij}$ positiv ist.

Diese Zusammenhänge sind in Abbildung 3.7 veranschaulicht:

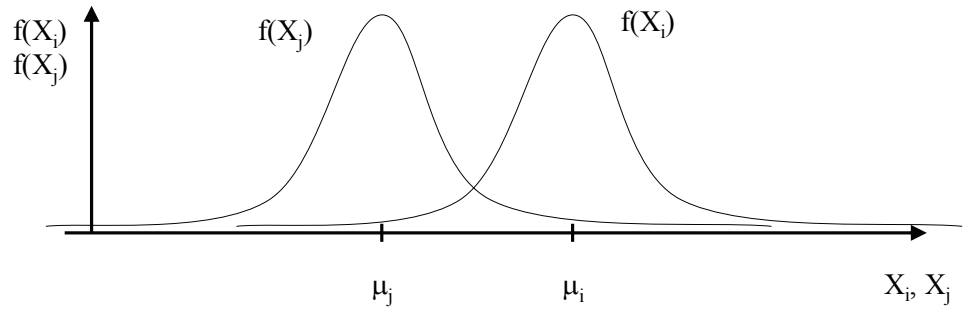


**Abbildung 3.7: Dichtefunktion der Bewertungen $X_i$ und $X_j$**

Die Erwartungswerte $\mu_i$ und $\mu_j$, die Standardabweichungen $\sigma_i$ und $\sigma_j$, sowie die Korrelation $\rho_{ij}$ sind wie bereits erwähnt unbekannt und müssen mit Hilfe des bekannten Anteils $p_{ij}$ geschätzt werden (vgl. GIERL, 1995, S. 610). Dies führt jedoch zu Problemen, da für die Schätzung der Parameter im Allgemeinen zu

wenige Beobachtungen vorliegen. Deshalb bedarf es einer weiteren Vereinfachung von Formel 3.4 (vgl. GIERL, 1995, S. 610). THURSTONE unterscheidet bei der Schätzung der Parameter mehrere Fälle, die zu einer Vereinfachung führen (1927, S. 273ff.). In seinem Fall 3 begründet er, weshalb die Annahme $\rho_{ij} = 0$ realistisch erscheint (THURSTONE, 1927, S. 279). Dadurch kann Formel 3.4 wie folgt vereinfacht werden (vgl. THURSTONE, 1927, S. 280):

$$\mu_i - \mu_j = z_{ij}\sqrt{\sigma_i^2 + \sigma_j^2} \tag{3.5}$$

Für die Schätzung der Parameter aus Formel 3.5 stehen verschiedene Verfahren zur Verfügung (vgl. GIERL, 1995, S. 610):

- THURSTONES Fall 5
- Verfahren von TORGERSON (basierend auf THURSTONES Fall 4)
- Maximum-Likelihood-Schätzverfahren

*THURSTONES Fall 5* stellt die leichteste Form der Parameterschätzung dar, da hier die stark vereinfachende Annahme getroffen wird, dass alle Standardabweichungen identisch sind (1927, S. 281). Diese Annahme wird beim *Verfahren von TORGERSON* dahin gehend abgeschwächt, dass die Varianzen lediglich ähnlich groß sind (1958, S. 164). Formal kann dies wie folgt ausgedrückt werden (vgl. TORGERSON, 1958, S. 164):

$$\mu_i - \mu_j = z_{ij}\sqrt{\sigma_i^2 + \sigma_i^2 + d} \tag{3.6}$$

*d:* $\sigma_j^2 - \sigma_i^2$

Die *Maximum-Likelihood-Schätzung* benötigt dagegen keine weiteren Annahmen (vgl. GIERL, 1995, S. 613). Die Schätzung der Parameter bei THURSTONES Fall 5 soll im Folgenden ausführlich geschildert werden. Für die beiden anderen Schätzverfahren sei auf die entsprechende Literatur verwiesen (siehe TORGERSON, 1958, S. 159-205 oder GIERL, 1995, S. 613-616).

Durch die Annahme identischer Standardabweichungen bei beiden Stimuli vereinfacht sich Formel 3.5 deutlich (vgl. THURSTONE, 1927, S. 282):

$$\mu_i - \mu_j = z_{ij}\sqrt{2\sigma^2} \tag{3.7}$$

Setzt man weiterhin voraus, dass die Erwartungswerte $\mu_i$ lediglich auf Intervallskalenniveau gemessen werden sollen, können zwei der drei Parameter beliebig gewählt werden (vgl. GIERL, 1995, S. 611). Es bietet sich an, Formel 3.7 durch Festlegung von $\sigma^2 = ½$ in folgenden Ausdruck zu überführen:

$$\mu_i - \mu_j = z_{ij} \qquad \Rightarrow \qquad \mu_i - \frac{1}{n}\sum_{j=1}^{n}\mu_j = \frac{1}{n}\sum_{j=1}^{n} z_{ij} \tag{3.8}$$

*n: Anzahl der Beurteilungsobjekte*

Die Schätzung der Erwartungswerte wird zusätzlich dadurch erleichtert, dass der Mittelwert aller Erwartungswerte $\mu$ gleich Null gesetzt werden kann (vgl. GIERL, 1995, S. 611). Damit ist folgende Schätzung für die Erwartungswerte $\mu_i$ möglich (vgl. GIERL, 1995, S. 611):

$$\mu_i = \frac{1}{n}\sum_{j=1}^{n} z_{ij} \tag{3.9}$$

Das Law of Comparative Judgement weist ähnliche Schwächen wie die Guttman-Skalierung auf. Probleme bereitet vor allem die Tatsache, dass in der Einstellungsforschung inhomogene Bewertungen der Stimuli nicht auf zufällige Schwankungen zurückzuführen sind, sondern durch unterschiedliche Präferenzen der Personen bedingt sind (vgl. GIERL, 1995, S. 54). Abhilfe könnte die mehrmalige Wiederholung des Paarvergleiches bei der gleichen Person schaffen, dies führt allerdings zu gravierenden erhebungstechnischen Schwierigkeiten (vgl. GIERL, 1995, S. 54). Wie bei allen eindimensionalen Skalierungsverfahren stößt das Law of Comparative Judgement außerdem an seine Grenzen, wenn das untersuchte Konstrukt mehrdimensional ist.

### 3.3.2.4 Die Unfolding-Technik

Die Unfolding-Technik von COOMBS ist eine Reaktion auf die empirisch zu beobachtenden Schwächen des Law of Comparative Judgement. Eine direkte Konsequenz dieses Modells ist folgende Transitivitätsüberlegung (vgl. COOMBS, 1983, S. 59):

$$p_{ij} \geq \frac{1}{2} \wedge p_{jk} \geq \frac{1}{2} \quad \Rightarrow \quad p_{ik} \geq max\left(p_{ij}, p_{jk}\right) \tag{3.10}$$

$p_{ij}$ beschreibt wiederum den empirisch beobachtbaren Anteil der Personen mit dem Urteil ($i \succ j$). Die Aussage $p_{ij} > ½$ ist ein Indiz dafür, dass der Erwartungswert von Stimulus i über dem Erwartungswert von Stimulus j liegt. Wenn dieser wiederum einen größeren Erwartungswert als Stimulus k aufweist, dann ist davon auszugehen, dass der Anteil $p_{ik}$ der Personen für die ($i \succ k$) gilt, mindestens so groß ist wie das Maximum aus $p_{ij}$ und $p_{jk}$. Dieser Zusammenhang wird jedoch in vielen Fällen empirisch widerlegt (vgl. COOMBS, 1983, S. 59).

Eine Erklärung für dieses Phänomen liefert die *Unfolding-Technik*. Diese besagt, dass Personen jene Objekte bevorzugen, die nahe an ihrem individuellen Idealpunkt (IP) liegen (vgl. KALLMANN, 1979, S. 89). Während beim Law of Comparative Judgement ($i \succ j$) durch $X_i > X_j$ bedingt ist, ist bei der Unfolding-Technik die Distanz zum Idealpunkt ($d_i < d_j$) entscheidend (siehe Abbildung 3.5):

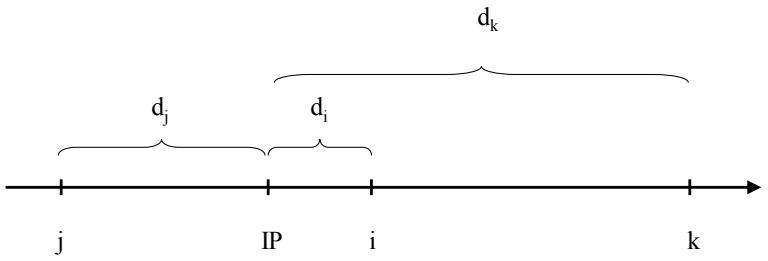


**Abbildung 3.8: Grundprinzip der Unfolding-Technik**

Aus den Distanzen zum Idealpunkt ergibt sich die Rangfolge der Objekte ($i \succ j \succ k$). Wie beim Law of Comparative Judgement ist auch bei der Unfolding-Technik die Lage der Objekte zufälligen Schwankungen unterworfen. Darüber hinaus unterliegt bei der Unfolding-Technik der Idealpunkt zufälligen Schwankungen (vgl. COOMBS, 1983, S. 60). Abbildung 3.9 verdeutlicht dies:

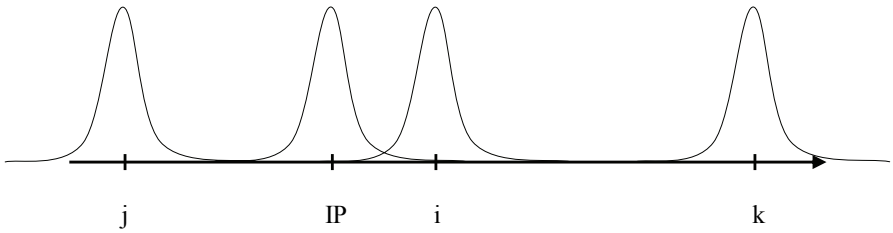


**Abbildung 3.9: Schwankungen bei der Objektbeurteilung**

In Abbildung 3.9 hat die Varianz des Idealpunktes keine Auswirkung auf die Präferenzbeurteilung des Objektpaares (i,k), da sich beide Objekte auf der selben Seite des Idealpunktes befinden. Für dieses Objektpaar sind lediglich die Standardabweichungen $\sigma_i$ und $\sigma_k$ relevant. Bei den Objektpaaren (i,j) und (j,k), welche auf unterschiedlichen Seiten des Idealpunktes liegen, ist dagegen zusätzlich die Standardabweichung des Idealpunktes ($\sigma_{IP}$) von Bedeutung. Durch eine Verschiebung des Idealpunktes in Richtung von j kann sich die Präferenz $j \succ i \succ k$ ergeben. Eine Verschiebung in Richtung k kann dagegen die Präferenz $i \succ k \succ j$ zur Folge haben. Für diese Objektpaare sind demnach größere Intransitivitäten zu erwarten (vgl. COOMBS, 1983, S. 61). Damit liefert die Unfolding-Technik eine Erklärung für die Verletzung der Transitivitätsüberlegung.

Kritisch anzumerken bleibt, dass die Unfolding-Technik zwar eine Erklärung, jedoch keine eindeutige Lösung zur Vermeidung der Transitivitätsverletzung liefert (vgl. KALLMANN, 1979, S. 90). Dennoch hat diese Technik eine große Bedeutung innerhalb der Skalierungsverfahren, da sie die Entwicklung der multidimensionalen Skalierung maßgeblich beeinflusst hat (vgl. WETTSCHUREK, 1974, S. 318).

### 3.3.3 Mehrdimensionale Skalierungsverfahren

Der grundlegende Nachteil aller eindimensionalen Skalierungsverfahren ist die zu starke Vereinfachung der Realität. So messen die bisher behandelten Verfahren nur eine Dimension eines interessierenden Konstrukts. Selbst wenn die einzelnen Items verschiedene Dimensionen eines Konstrukts erfragen, bleibt das Problem bestehen, dass ein einziger Gesamtpunktwert ohne Berücksichtigung der Dimensionen berechnet wird (vgl. KALLMANN, 1979, S. 94). Dies entspricht nicht dem mehrdimensionalen Charakter vieler Konstrukte.

Diesen Nachteil versuchen mehrdimensionale Skalierungsverfahren zu vermeiden. Ein Teil der Verfahren versucht die Mehrdimensionalität durch geeignete Verknüpfungsregeln abzubilden. Das Ergebnis dieser Verfahren ist ein einziger Wert. Andere Verfahren gehen noch einen Schritt weiter und tragen der Mehrdimensionalität dadurch Rechnung, dass sie ein Zahlentupel als Ergebnis liefern.

Jedes dieser Verfahren setzt sich aus der Kombination verschiedener Ratingskalen oder Paarvergleichen zusammen. Die in Kapitel 3.3.1 angesprochenen grundlegenden Verfahren sind also erneut relevant.

In Abbildung 3.10 sind die verschiedenen Verfahren dahingehend strukturiert, ob sie einen einzelnen Wert oder ein Zahlentupel ergeben und auf welches grundlegende Verfahren sie dabei zurückgreifen:

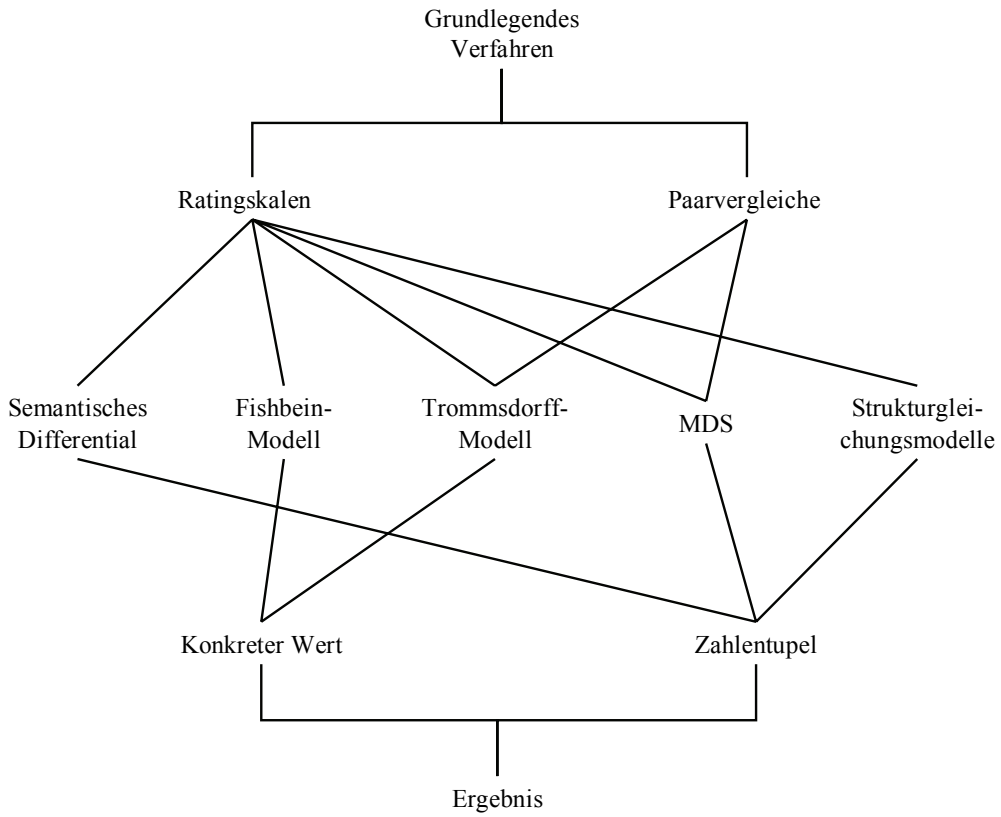


**Abbildung 3.10: Einteilung der mehrdimensionalen Skalierungsverfahren**

Das semantische Differential unternimmt nicht den Versuch, einzelne Indikatoren zu aggregieren. Daher liefert es auch keinen konkreten Wert als Ergebnis. Alle übrigen Verfahren fassen einzelne Indikatoren zusammen. Bei den Modellen von FISHBEIN und TROMMSDORFF führt diese Aggregation zu einem bestimmten Wert, bei der multidimensionalen Skalierung (MDS) und Strukturgleichungsmodellen unter Berücksichtigung der Dimensionen zu einem Zahlentupel.

Daher erscheint folgende Strukturierung des Kapitels sinnvoll:

- Kapitel 3.3.3.1: Das semantische Differential von OSGOOD, SUCI und TANNENBAUM
- Kapitel 3.3.3.2: Das Modell von FISHBEIN
- Kapitel 3.3.3.3: Das Modell von TROMMSDORFF
- Kapitel 3.3.3.4: Die multidimensionale Skalierung (MDS)
- Kapitel 3.3.3.5: Strukturgleichungsmodelle

### 3.3.3.1 Das semantische Differential

Das *semantische Differential* misst die Assoziation, die eine Versuchsperson mit einem Objekt verbindet (vgl. MEFFERT, 1992, S. 186). Dabei sind jedoch keine freien Assoziationen zulässig, sondern lediglich die Bewertung für vorher festgelegte Begriffspaare (vgl. HAMMANN, ERICHSON, 2000, S. 281). Die Begriffe sollten für ein semantisches Differential abstrakt formuliert sein (vgl. GIERL, 1995, S. 61). Um die Stärke einer Assoziation messen zu können, kommen bipolare Ratingskalen zum Einsatz (vgl. HAMMANN, ERISCHON, 2000, S. 281). Die einzelnen Antwortkategorien sind für jedes Objektpaar semantisch abgestuft (vgl. HAMMANN, ERISCHON, 2000, S. 281). Die Anzahl der Antwortmöglichkeiten variiert, häufig sind sieben Kategorien vorgegeben (vgl. MEFFERT, 1992, S. 186). Im Gegensatz zur eindimensionalen Skalierung erfolgt nicht der Versuch, die Begriffspaare zu aggregieren (vgl. HAMMANN, ERISCHON, 2000, S. 281). Stattdessen stellt man für jedes Objekt ein Einstellungsprofil auf (vgl. MEFFERT, 1992, S. 186). In Abbildung 3.11 ist die Bewertung einer Versuchsperson für zwei verschiedene Beurteilungsobjekte angegeben. Die Mehrdimensionalität entsteht durch die verwendeten Begriffspaare, die verschiedene Dimensionen des Konstrukts abdecken sollten.

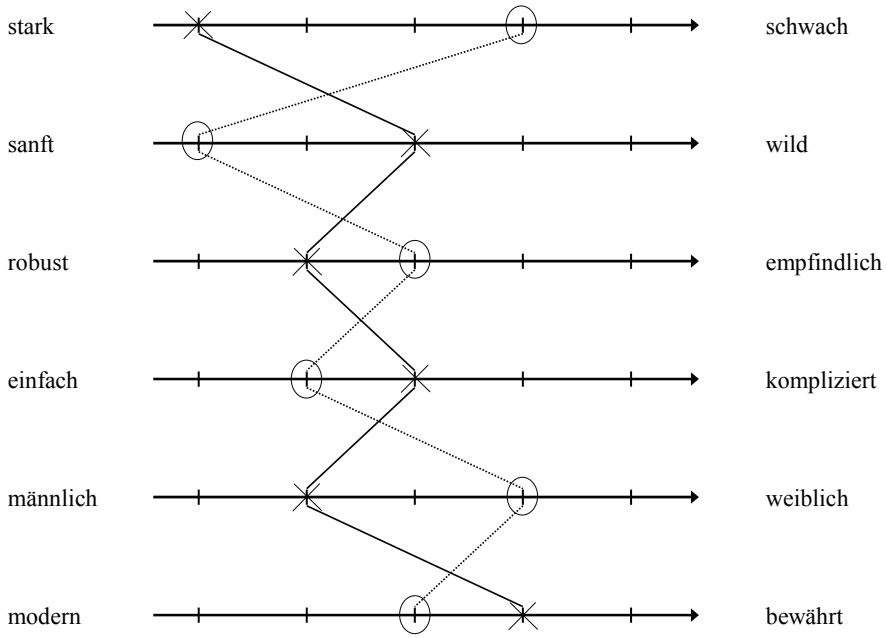


**Abbildung 3.11: Beispiel für ein semantisches Differential (in Anlehnung an KROEBER-RIEL, WEINBERG, 1999, S. 198)**

Die Gegenüberstellung der durch ein semantisches Differential erstellten Einschätzungsprofile für verschiedene Objekte liefert Erkenntnisse über Gemeinsamkeiten und Besonderheiten. In Abbildung 3.11 könnte beispielsweise die Einschätzung einer Person zu verschiedenen PKW-Marken erhoben sein. Bei einer Befragung mehrerer Personen kann auch eine Mittelwertbildung über alle Personen erfolgen und dieser Wert abgetragen werden.

Ein großer Nachteil des semantischen Differentials ist der große Interpretationsspielraum. So geht aus den Ergebnissen aus Abbildung 3.11 nicht hervor, ob ein Proband die Ausprägung sanft oder wild bevorzugt. Eine Bewertung der Einstellung kann daher nicht vorgenommen werden. Es können deshalb lediglich Unterschiede zwischen verschiedenen Untersuchungsobjekten verdeutlicht werden, ohne die tatsächliche Einstellung zu messen.

### 3.3.3.2 Das Modell von FISHBEIN

Das Modell von FISHBEIN versucht auf Basis der Messung affektiver und kognitiver Komponenten, eine gemeinsame Kennzahl für Konstrukte (insbesondere die Einstellung) zu ermitteln (1963, S. 233ff.). Damit ist dieses Modell prinzipiell

ein eindimensionales Skalierungsverfahren, das die Mehrdimensionalität von Konstrukten nur unzureichend wiedergeben kann. Dennoch soll es den mehrdimensionalen Skalierungsverfahren zugeordnet werden, da es bei der Aggregation der Indikatoren den Versuch unternimmt, die Mehrdimensionalität zu berücksichtigen. Dazu sind zunächst zwei Voraussetzungen relevant (vgl. HAMMANN, ERICHSON, 2000, S. 283):

1. Jeder Stimulus besitzt Merkmale bzw. Eigenschaften, welche die Einstellung der Testpersonen vorrangig beeinflussen.

2. Die Einstellung zu einem Objekt setzt sich aus subjektivem Wissen der Testpersonen um diese Eigenschaften und deren subjektiver Bewertung zusammen.

Zwei auf ROSENBERG zurückgehende Annahmen ermöglichen die Operationalisierung (1967, S. 325ff.). Die wahrgenommenen Eigenschaften bestehen vor allem aus kognitiven Komponenten, die bewerteten Eigenschaften dagegen eher aus affektiven Komponenten. Die *Multiplikativitätsprämisse* setzt voraus, dass diese beiden Dimensionen multiplikativ verknüpft sind. Die *Additivitätsprämisse* besagt, dass die Einstellung eine Linearkombination aller wahrgenommenen (erste Dimension) und bewerteten (zweite Dimension) Eigenschaften ist.

Für jede Eigenschaft muss somit geprüft werden, wie hoch die Möglichkeit von der Versuchsperson eingeschätzt wird, dass eine bestimmte Ausprägung beim Beurteilungsobjekt vorhanden ist (erste Dimension). Ein Beispiel ist die Frage: ´Wie wahrscheinlich ist es, dass Autos der Marke X sicher sind?´ Auf Grund des Wertebereichs [0;1] wird dieser Wert in der Literatur meist als Wahrscheinlichkeit bezeichnet.

Zur Bewertung (zweite Dimension) gelangt man, indem die Personen auf einer Ratingskala angeben, wie positiv das Vorhandensein dieser Eigenschaft für sie ist (vgl. HAMMANN, ERICHSON, 2000, S. 284). Im angeführten Beispiel entspricht dies der Frage, wie positiv eine hohe Sicherheit bei einem PKW ist.

$$R_{ijk} = p_{ijk} Q_{ik} \qquad \textit{(Multiplikativitätsprämisse)} \tag{3.11}$$

$R_{ijk}$: *Eindruckswert der Eigenschaft k von Objekt j bei Person i*

$p_{ijk}$: *Überzeugungsgrad von Person i, dass Eigenschaft k bei j vorhanden ist*

$Q_{ik}$: *Bewertung der Eigenschaft k durch Person i*

Die beiden Dimensionen werden multiplikativ verknüpft und ergeben einen Eindruckswert für diese Eigenschaft. Der Einstellungswert zu einem bestimmten Objekt ergibt sich durch Addition aller Eindruckswerte (vgl. KALLMANN, 1979, S. 87).

$$E_{ij} = \sum_{k=1}^{m} R_{ijk} \qquad \textit{(Additivitätsprämisse)} \tag{3.12}$$

$E_{ij}$: *Einstellungswert der Person i zum Stimulus j*

Ein Nachteil der Methode von FISHBEIN ist, dass die Eindruckswerte ungewichtet in die Berechnung des Einstellungswertes einfließen (vgl. HAMMANN, ERICHSON, 2000, S. 285). Die Personen nehmen zwar eine Einschätzung des Überzeugungsgrads für das Vorhandensein einer Ausprägung vor, jedoch keine Gewichtung in Bezug auf die Bedeutung der Eigenschaft (vgl. HAMMANN, ERICHSON, 2000, S. 285). Diese könnte im Rahmen einer Erweiterung des Modells zusätzlich erfragt werden. Ob die Probanden allerdings in der Lage sind, die Bedeutung der einzelnen Merkmale korrekt anzugeben, erscheint fraglich.

Es sind viele Konstrukte vorstellbar, bei denen die Prämissen über die Verknüpfung der Dimensionen unrealistisch erscheinen. Insofern ist das Modell von FISHBEIN kein allgemein geeignetes Modell zur Messung mehrdimensionaler Konstrukte und deshalb nur ansatzweise ein mehrdimensionales Skalierungsverfahren. Falls die Prämissen jedoch gerechtfertigt erscheinen, können mit ihm auch mehrdimensionale Konstrukte gemessen werden.

### 3.3.3.3 Das Modell von TROMMSDORFF

Analog zum Modell von FISHBEIN besteht beim Modell von TROMMSDORFF die wahrgenommene Merkmalsausprägung vor allem aus kognitiven Komponenten und die Bewertung der Merkmalsausprägung aus affektiven Komponenten. Die wahrgenommene Merkmalsausprägung wird direkt erfragt, während die Bewertung der Merkmalsausprägungen nur indirekt, mit Hilfe der Distanz zum Idealpunkt, bestimmt wird (vgl. TROMMSDORFF, 1989, S. 67ff.). Zur Bestimmung dieser Distanz ist die Information, welche Merkmalsausprägungen der Proband als ideal erachtet, erforderlich. Damit ist dieses Modell eine Kombination aus Unfolding-Technik und dem Modell von FISHBEIN. Die Versuchsperson gibt auf einer Ratingskala wahrgenommene Ausprägungen realer

Untersuchungsobjekte (kognitive Komponenten) und subjektive ideale Ausprägungen (affektive Komponenten) an (vgl. HAMMANN, ERICHSON, 2000, S. 287).

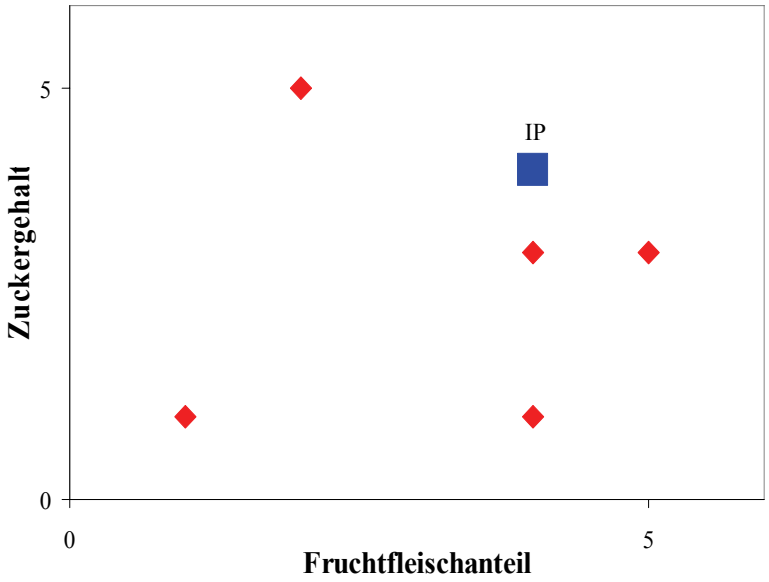


**Abbildung 3.12: Beispiel für eine Skalierung mit dem Modell von TROMMSDORFF**

Abbildung 3.12 veranschaulicht die Idee des Verfahrens von TROMMSDORFF, wenn eine Person ihren subjektiven Eindruck des Fruchtfleischanteils und Zuckergehalts von fünf Orangensaftmarken angibt. Mit IP sind die von der Person als ideal angegebenen Merkmalsausprägungen in die Repräsentation integriert. Der Eindruckswert basiert auf der Distanz der wahrgenommenen Merkmalsausprägung zum Idealpunkt.

$$R_{ijk} = \left| A_{ijk} - I_{ik} \right| \tag{3.13}$$

$R_{ijk}$: *Eindruckswert der Eigenschaft k von Objekt j bei Person i*

$A_{ijk}$: *Von Person i wahrgenommene Ausprägung des Merkmals k bei j*

$I_{ik}$: *Von Person i empfundene ideale Ausprägung des Merkmals k*

Der Einstellungswert zu einem Objekt kann dann analog zum Verfahren von FISHBEIN durch Addition der Eindruckswerte gewonnen werden (vgl. HAMMANN, ERICHSON, 2000, S. 286). Für die Berechnung der Distanz zum

Idealpunkt wären auch andere Distanzmaße, wie sie im Rahmen der multidimensionalen Skalierung behandelt werden, denkbar.

$$E_{ij} = \sum_{k=1}^{m} R_{ijk} \qquad (3.14)$$

$E_{ij}$: *Einstellungswert der Person i zu Stimulus j*

Problematisch am Modell von TROMMSDORFF ist die kompositionelle Messung der idealen Ausprägung (vgl. HAMMANN, ERICHSON, 2000, S. 291). Bei vielen Fragestellungen sind die Probanden mit der Angabe optimaler Produkteigenschaften nämlich überfordert. Es ist zum Beispiel nicht anzunehmen, dass eine Person einer Brauerei mitteilen kann, welchen Alkoholgehalt, Resteextraktanteil und ähnliches ein ideales Bier aufweisen muss. Darüber hinaus kann die ungewichtete Berechnung der merkmalsweisen Distanzen zum Idealpunkt kritisch sein, wenn die Wertebereiche der Variablen unterschiedlich sind. In diesem Fall dominieren vor allem die Merkmale mit stark variierenden Wertebereichen. Eine Standardisierung der Merkmale könnte dieses Problem verhindern. Ob die Mehrdimensionalität von Konstrukten allein durch die Unterscheidung von wahrgenommener und als ideal empfundener Ausprägung ausreichend berücksichtigt ist, scheint außerdem sehr fraglich. Vielmehr versucht das Modell von TROMMSDORFF einen einzigen Skalenwert für die Objekte zu berechnen. Insofern stellt es keine echte mehrdimensionale Methode dar.

### 3.3.3.4 Multidimensionale Skalierung

Bei der multidimensionalen Skalierung (MDS) sind in der Literatur mehrere Verfahrensvarianten diskutiert. In dieser Arbeit soll lediglich die MDS nach KRUSKAL dargestellt werden (vgl. KRUSKAL, 1964, S. 115-129). Die Grundidee von KRUSKAL ist allerdings auch auf andere Varianten, etwa die metrische MDS nach SHEPARD, übertragbar (vgl. SHEPARD, 1962, S. 125-139). Das Ziel der multidimensionalen Skalierung ist die grafische Repräsentation der Ähnlichkeitsbeziehungen zwischen n Objekten in einem q-dimensionalen Raum mit minimalem Informationsverlust (vgl. HAIR et al., 2006, S. 629). Aus Übersichtlichkeitsgründen wird häufig q = 2 gewählt (vgl. GIERL, 1995, S. 122). Damit verfolgt die MDS ein ähnliches Ziel wie das Modell von TROMMSDORFF oder das semantische Differential (vgl. HAMMANN, ERICHSON, 2000, S. 290). Im

Gegensatz zu diesen beiden Verfahren beruht die MDS allerdings auf direkt erhobenen Ähnlichkeitsdaten (vgl. BACKHAUS et al., 2006, S. 621). Dafür stehen die Paarvergleichsmethode, die Ankerpunktmethode oder die Einschätzung der Ähnlichkeit zweier Objekte mittels einer Ratingskala zur Verfügung (vgl. HAIR et al., 2006, S. 635). Die Ankerpunktmethode ist eine spezielle Form des Paarvergleichs (vgl. BACKHAUS et al., 2006, S. 628f.) Die Ermittlung der Ähnlichkeitsdaten durch den Vergleich von bewerteten Merkmalsausprägungen ist ebenfalls denkbar, jedoch weisen Paarvergleiche und Ratingskalen den Vorteil auf, dass den Versuchspersonen die Merkmalsausprägungen der Objekte nicht bekannt sein müssen und auch die individuelle Gewichtung der Merkmale nicht benötigt wird (vgl. HAMMANN, ERICHSON, 2000, S. 290). Des Weiteren kann bei direkt befragten Wahrnehmungen die interne Konsistenz nicht überprüft werden (vgl. GIERL, 1995, S. 116). Bei Ähnlichkeitsurteilen ist dies durch Überprüfung der Transitivität möglich.

Die so erhobenen Daten spiegeln die von einer Person empfundenen Ähnlichkeitsbeziehungen der Objekte wider. Bei der Bestimmung der Ähnlichkeit zweier Objekte geht man meist zur Bildung von Distanzen über, da bei einem Distanzmaß die erhaltenen Werte einfacher interpretierbar sind (vgl. HANDL, 2002, S. 83). Zwei von der Versuchsperson identisch wahrgenommene Objekte ($i \approx j$) erhalten die Distanz Null. Je unähnlicher die Person zwei Objekte einstuft, desto größer wird dieser Wert. Folgende Eigenschaften können für Distanzmaße gefordert werden (vgl. OPITZ, 1980, S. 30):

$$\begin{aligned} &d(i,i) = 0 && \textit{(Reflexivität)} \\ &d(i,j) = d(j,i) && \textit{(Symmetrie)} \\ &d(i,j) \geq 0 && \textit{(Nichtnegativität)} \end{aligned} \tag{3.15}$$

Weitere in der Literatur genannte Voraussetzungen an die Distanzen, etwa die Dreiecksungleichung, sind für die MDS nach KRUSKAL nicht erforderlich (vgl. OPITZ, 1980, S. 126). Fasst man die Distanzen zwischen den Objekten zu einer Distanzmatrix D zusammen, so muss diese unter Berücksichtigung von (3.15) quadratisch, symmetrisch und auf der Hauptdiagonalen ausschließlich mit Nullen besetzt sein.

Eine Repräsentation $X = (x_{ik})_{nq}$ enthält in den n Zeilen die zu repräsentierenden Objekte und in den q Spalten die Koordinaten einer bestimmten Dimension (vgl.

OPITZ, 1980, S. 126). Für diese Repräsentation sind die Distanzen $\hat{d}$ der n Objekte berechenbar (vgl. BACKHAUS et al., 2006, S. 633):

$$\hat{d}(i,j) = \left( \sum_{k=1}^{q} \left| x_{ik} - x_{jk} \right|^{p} \right)^{\frac{1}{p}} \tag{3.16}$$

$x_{ik}$: *Koordinate von Objekt i bezüglich Dimension k*

Wählt man den Parameter p = 1, erhält man die City-Block-Metrik, für p = 2 entspricht das Maß der euklidischen Distanz (vgl. BACKHAUS et al., 2006, S. 631f.). Für die Distanzberechnung einer zweidimensionalen Repräsentation erscheint die Verwendung der euklidischen Distanz deshalb sinnvoll, da diese die direkte Verbindung zweier Punkte angibt (vgl. HANDL, 2002, S. 85). Die Repräsentation X ist perfekt, wenn folgende Monotoniebedingung stets zutrifft (vgl. BACKHAUS et al., 2006, S. 634):

$$d(i,j) \leq d(i',j') \quad \Rightarrow \quad \hat{d}(i,j) \leq \hat{d}(i',j') \tag{3.17}$$

Bei Verletzung der Monotoniebedingung ist ein Gütemaß zu bestimmen, das eine Entscheidung über die Qualität der Repräsentation ermöglicht (BACKHAUS et al., 2006, S. 636). Als Gütemaß für eine vorliegende Repräsentation gilt die Abweichung von einer perfekten, die Monotoniebedingung erfüllenden Lösung. Die Repräsentation ist umso schlechter, je mehr sie von der perfekten Lösung abweicht. Die Abweichung von der perfekten Lösung kann durch Berechnung der Disparitäten $\delta$ bestimmt werden. Die Disparitäten sind, ähnlich wie beim Verfahren der monotonen Regression, an die Ausgangsdistanzmatrix D monoton angepasste Zahlen (vgl. OPITZ, 1980, S. 131). Das heißt, die Disparitäten werden so bestimmt, dass sie möglichst nahe an den Werten $\hat{d}$ liegen, aber dennoch die Monotoniebedingung erfüllen (vgl. HANDL, 2002, S. 162):

$$d(i,j) \leq d(i',j') \quad \Rightarrow \quad \delta(i,j) \leq \delta(i',j') \tag{3.18}$$

Zur Berechnung der Disparitäten ordnet man die Objektpaare aufsteigend nach ihren ursprünglichen Distanzen und vergleicht diese mit den berechneten Distanzen (vgl. HANDL, 2002, S. 162). Falls die Monotoniebedingung 3.17 erfüllt ist, entsprechen die Disparitäten den berechneten Distanzen:

$$\delta(i,j)_r = \hat{d}(i,j)_r \tag{3.19}$$

Der Parameter r beschreibt den Platz des Objektpaares (i,j) in der Rangordnung der ursprünglichen Distanzen. Wird die Monotoniebedingung (3.17) bei zwei oder mehr aufeinander folgenden Objektpaaren nicht eingehalten, errechnet man die Disparitäten durch Mittelwertbildung der berechneten Distanzen, in (3.20) für zwei Objektpaare dargestellt (vgl. BACKHAUS et al., 2006, S. 637):

$$\delta(i,j)_r = \delta(i',j')_{r+1} = \frac{1}{2}\left(\hat{d}(i,j)_r + \hat{d}(i',j')_{r+1}\right) \tag{3.20}$$

Da die Mittelwertbildung nicht bei allen Paaren vorgenommen wird, kann nach dieser Berechnung jedoch die Monotoniebedingung (3.18) wegen $\delta(i,j)_{r-1} > \delta(i,j)_r$ verletzt sein (vgl. OPITZ, 1980, S. 132). Daher sind im Anschluss an die Mittelwertbildung wieder alle Objektpaare von Beginn an auf die Erfüllung von (3.18) zu überprüfen (vgl. HANDL, 2002, S. 162). Nachdem mit dieser Vorgehensweise die Gültigkeit von (3.18) für alle Objektpaare sichergestellt ist, kann die Güte der Repräsentation bestimmt werden. Sie entspricht dann der Summe der quadrierten Abweichung der Disparitäten von den berechneten Distanzen (vgl. BACKHAUS et al., 2006, S. 639). Dieser Wert wird *Stress* (b) der Repräsentation genannt:

$$b = \sum_{i<j}\left(\hat{d}(i,j) - \delta(i,j)\right)^2 \tag{3.21}$$

Bei einer perfekten Anpassung ist der Stress gleich Null (vgl. BACKHAUS et al., 2006, S. 639). Kommt es zu einer Verletzung der Monotoniebedingung, steigt der Wert an. Da dieser Anstieg auch von der Streuung der berechneten Distanzen $\hat{d}$ abhängt, können keine Vergleiche verschiedener Repräsentationen vorgenommen werden (vgl. HANDL, 2002, S. 163). Daher normiert man den Stress durch die Division mit dem maximal möglichen Wert für eine Repräsentation auf den Wertebereich [0;1] (vgl. BACKHAUS et al., 2006, S. 639). Der maximal mögliche Stress für eine Repräsentation entsteht, wenn (3.17) stets verletzt ist:

$$d(i,j) \leq d(i',j') \quad \Rightarrow \quad \hat{d}(i,j) > \hat{d}(i',j') \tag{3.22}$$

In diesem Fall wird zur Bestimmung der Disparitäten der Mittelwert aller n(n-1)/2 berechneter Distanzen der Objektpaare gebildet. Deshalb kann eine Obergrenze angegeben werden, die der Stress einer Repräsentation nicht überschreiten kann (vgl. HAIR et al., 2006, S. 653):

$$b_{max} = \sum_{i<j} \left( \hat{d}(i,j) - \bar{\hat{d}} \right)^2 \tag{3.23}$$

Dieser Wert wird als Normierungsfaktor für (3.21) verwendet. Der Quotient dieser Werte ergibt den normierten Stress der Funktion, der kleiner als 0,2 sein sollte (vgl. HAIR et al., 2006, S. 653).

$$b_{norm} = \frac{b}{b_{max}} \tag{3.24}$$

Beispiel 3.2 veranschaulicht die Berechnung des Stress-Wertes:

**Beispiel 3.2**

*Für fünf Objekte seien folgende Distanzbeziehungen gegeben:*

*d(1,2) < d(1,3) < d(2,3) < d(4,5) < d(1,4) < d(1,5) < d(2,4) < d(3,4) < d(2,5) < d(3,5)*

*Für eine zweidimensionale Repräsentation X erhalte man folgende Daten:*

$$\begin{pmatrix} -2 & -1 \\ -1 & -1 \\ -3 & -3 \\ 0 & 2 \\ 3 & 4 \end{pmatrix}$$

*Durch die Anordnung der Objektpaare nach wachsenden Distanzen d(i,j) (siehe Tabelle 3.4) erkennt man, dass die berechneten euklidischen Distanzen lediglich beim Vergleich der Objektpaare (1,5) und (2,4) die Monotoniebedingung verletzen. Die Disparitäten können daher wie folgt bestimmt werden:*

| (i,j) | (1,2) | (1,3) | (2,3) | (4,5) | (1,4) | (1,5) | (2,4) | (3,4) | (2,5) | (3,5) |
|---|---|---|---|---|---|---|---|---|---|---|
| $\hat{d}(i,j)$ | 1 | 2,2 | 2,8 | 3,6 | 3,6 | 7,1 | 3,2 | 5,8 | 6,4 | 9,2 |
| $\delta(i,j)$ | 1 | 2,2 | 2,8 | 3,6 | 3,6 | 5,2 | 5,2 | 5,8 | 6,4 | 9,2 |

**Tabelle 3.4: Berechnung der Disparitäten**

*Für diese Repräsentation kann ein Stress von 7,61 berechnet werden. Da der maximal mögliche Stress 57,889 beträgt, ergibt sich folgender normierter Stress:*

$$b_{norm} = \frac{7,61}{57,889} \approx 0,13$$

Nachdem ein Maß zur Bewertung einer Repräsentation vorhanden ist, muss als nächstes eine möglichst gute Repräsentation gefunden werden. Dies geschieht durch kontinuierliche Verbesserung einer zufälligen Startlösung $X^0$. Die MDS ist somit ein iteratives Verfahren, das abbricht, wenn eine gefundene Repräsentation nicht mehr oder nur noch minimal verbessert werden kann (vgl. BACKHAUS et al., 2006, S. 640). Das Ziel der Verbesserung ist eine Minimierung des Stresses. Daher bildet man ausgehend von einer zufälligen Startlösung $X^0$ den negativen Gradienten des unnormierten Stress-Wertes (vgl. OPITZ, 1980, S. 133):

$$-B = -\left(\frac{\partial b}{\partial x_{ik}}\right)_{n,q} \tag{3.25}$$

Mit Hilfe eines Schrittweitenparameters $\lambda$ berechnet sich die Folgerepräsentation $X^1$ wie in (3.26) beschrieben. Die Größe für die Schrittweite kann dabei frei festgelegt werden. Erfahrungsgemäß ist der Startwert 0,2 für $\lambda_0$ günstig, der im Fortlauf des Algorithmuses sinkt (vgl. BACKHAUS et al., 2006, S. 641).

$$X^1 = X^0 - \lambda_0 B \tag{3.26}$$

In Abbildung 3.13 ist eine zweidimensionale Repräsentation mit Stress 0 für Beispiel 3.2 angegeben. Zusätzlich ist auch die Startlösung $X^0$ enthalten:

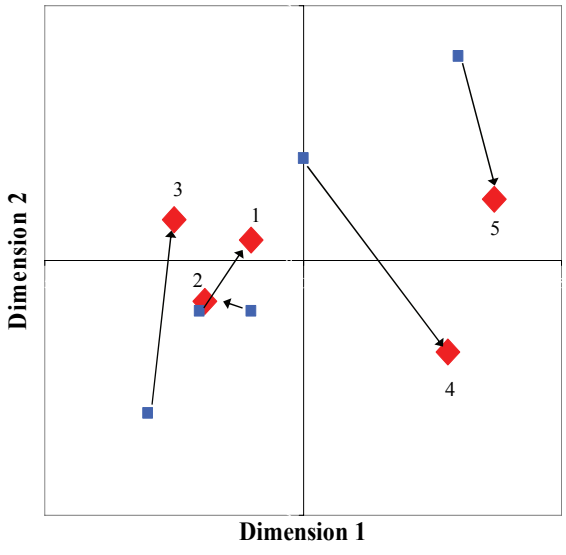


**Abbildung 3.13: Ergebnis einer zweidimensionalen MDS für Beispiel 3.2**

Obwohl die MDS die Mehrdimensionalität von Konstrukten berücksichtigt, indem alle Dimensionen in die Ähnlichkeitsbeurteilung der Probanden einfließen können und diese zusätzlich keine Kenntnis einzelner Merkmalsausprägungen benötigen, weist die MDS auch folgende Nachteile auf (vgl. HAMMANN, ERICHSON, 2000, S. 302):

- Festlegung der Anzahl der Dimensionen (q) nicht verfahrensimmanent
- Keine Berücksichtigung individueller Unterschiede in der Wahrnehmung
- Schwierigkeiten bei der Dimensionsinterpretation
- Keine Bewertung der Einstellung der Person
- Überforderung der Testpersonen bei Bestimmung der Ähnlichkeitsdaten

Die *Anzahl der Dimensionen* sollte eigentlich den tatsächlich vorhandenen Dimensionen des Konstrukts entsprechen. Auf Grund der begrenzten räumlichen Vorstellungskraft reduziert sich das Problem jedoch auf die Frage, ob zwei oder drei Dimensionen sinnvoll sind (vgl. BACKHAUS et al., 2006, S. 645). Damit ist die Festlegung der Dimensionen eher unproblematisch. Falls das zu untersuchende Konstrukt jedoch mehr als drei Dimensionen aufweist, können diese unter Umständen nicht adäquat repräsentiert werden.

Die *Berücksichtigung individueller Unterschiede* kann durch drei unterschiedliche Herangehensweisen erfolgen (vgl. HAIR et al., 2006, S. 640).

Die einfachste Methode besteht in der Mittelwertbildung aller Distanzen (vgl. BACKHAUS et al., 2006, S. 647). Falls die Einstellungen der Personen jedoch sehr heterogen sind, ist diese Vorgehensweise kritisch. Es empfiehlt sich dann, die Personen durch eine Clusteranalyse zu homogenen Segmenten, bezogen auf die Distanzen, zusammenzufassen und die MDS für jedes Cluster anzufertigen (vgl. HAIR et al., 2006, S. 641).

Alternativ kann auch für jede Person eine individuelle Repräsentation berechnet werden und die Position der Objekte wird durch Mittelwertbildung aggregiert (vgl. BACKHAUS et al., 2006, S. 647). Bei sehr unterschiedlichen Repräsentationen empfiehlt sich auch hier die Clusteranalyse zur Bestimmung homogener Segmente. In diesem Fall sind die Segmente jedoch bezüglich der Koordinaten der Objekte homogen (vgl. HAIR et al., 2006, S. 641).

Es existieren aber auch Software-Pakete wie ALSCAL, die eine gemeinsame Analyse der Ähnlichkeitsdaten zulassen und eine gemeinsame Konfiguration ermitteln (vgl. BACKHAUS et al, 2006, S. 647).

Die Schwierigkeiten bei der *Dimensionsinterpretation* löst das Property-Fitting. Dazu müssen die wahren Ausprägungen der Objekte bei ordinalen oder metrischen Merkmalen bekannt sein oder im Rahmen der Befragung zusätzlich erhoben werden (vgl. BACKHAUS et al., 2006, S. 669). Diese können dann mit Hilfe linearer oder monotoner Regression eingebettet werden, wobei die Merkmalsausprägung als abhängige Variable und die Koordinaten der Objekte als unabhängige, erklärende Merkmale in die Regressionsgleichung einfließen (vgl. BACKHAUS et al., 2006, S. 659-669).

Abbildung 3.14 zeigt ein Beispiel für einen eingebetteten Merkmalsvektor. Dem Vektor ist folgende Beziehung der Objekte, bezogen auf das eingebettete Merkmal, zu Grunde gelegt: $O_5 > O_4 > O_1 > O_2 > O_3$

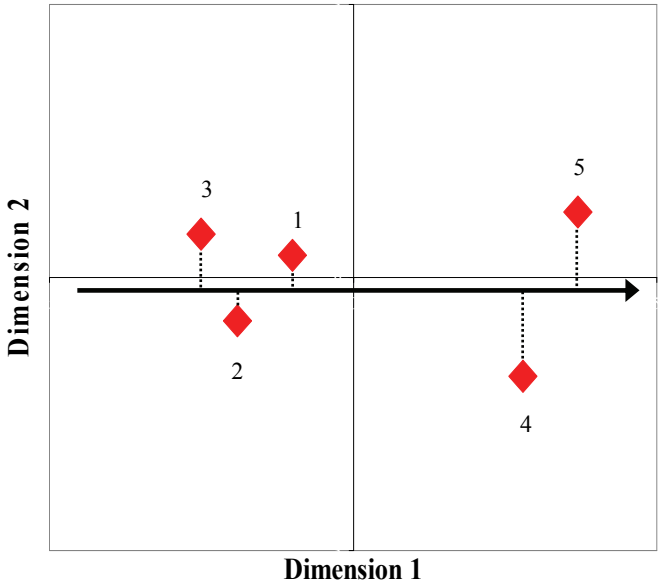


**Abbildung 3.14: Veranschaulichung der Merkmalseinbettung bei der MDS**

Mit Hilfe des Property-Fitting können beliebige Merkmale, ausgenommen nominal-polytome, in die MDS integriert werden. Diese Merkmalsvektoren helfen bei der Interpretation der Achsen.

Die Lage der Objekte im zweidimensionalen Raum und die Interpretation der Achsen ermöglicht aber noch keine Aussage, welche Objekte von der Auskunftsperson als besonders positiv eingestuft werden. Da diese Fragestellung im Rahmen der Einstellungsmessung von großem Interesse ist, müssen die Personen zusätzlich ihre Präferenz für die Objekte angeben (vgl. GIERL, 1995, S. 130). Dies kann mit Hilfe einer Rangordnung oder einer Ratingskala geschehen (vgl. BACKHAUS et al., 2006, S. 654). Durch diese Präferenz kann dann ein Idealpunkt oder -vektor in die Repräsentation so eingefügt werden, dass die angegebene Präferenz durch die Distanz der Objekte zum Idealpunkt bzw. der Lage auf dem Idealvektor bestmöglich wiedergegeben wird (vgl. GIERL, S. 130f.). Als Instrument zur Einpassung dient wie beim Property-Fitting ein linearer oder monotoner Regressionsansatz (vgl. BACKHAUS et al., 2006, S. 659).

Möchte man einen Idealpunkt für eine Person bestimmen, deren Präferenzen mit Hilfe einer Ratingskala ermittelt wurden, so erhält man bei einer zweidimensionalen Repräsentation nach Bestimmung der Regressionskoeffizienten beispielsweise folgenden Schätzwert für Objekt i (vgl. HAMMANN, ERICHSON, 2000, S. 322):

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i1}^{\,2} + \hat{\beta}_4 x_{i2}^{\,2} \qquad (3.27)$$

$\hat{y}_i$ : *Geschätzter Punktwert für Objekt i*

$x_{i1}$: *Koordinate von Objekt i bezüglich Dimension 1*

$x_{i2}$: *Koordinate von Objekt i bezüglich Dimension 2*

Der geschätzte Punktwert soll möglichst exakt den vom Probanden angegebenen Punktwert y wiedergeben. Da im Allgemeinen die beiden Dimensionen als gleichbedeutend angesehen werden, kann auch $\beta_3 = \beta_4$ gesetzt werden. Die Regressionskoeffizienten geben Auskunft über die Lage des Idealpunktes (vgl. HAMMANN, ERICHSON, 2000, S. 322):

$$IP = \left( -\frac{\hat{\beta}_1}{2\hat{\beta}_3} \; ; \; -\frac{\hat{\beta}_2}{2\hat{\beta}_4} \right) \qquad (3.28)$$

Eine Person kann in Beispiel 3.2 etwa folgende Objektbewertungen auf einer Ratingskala angeben (höhere Punktzahlen bedeuten eine positivere Bewertung des Objekts):

$O_5$: 5 $O_1$: 4 $O_4$: 3 $O_3$: 2 $O_2$: 1

Mit Hilfe der linearen Regression und $\beta_3 = \beta_4$ erhält man für den Idealpunkt die Koordinaten (1,8 ; 1,9). In Abbildung 3.15 ist die Einbettung des Idealpunktes in die Repräsentation veranschaulicht:

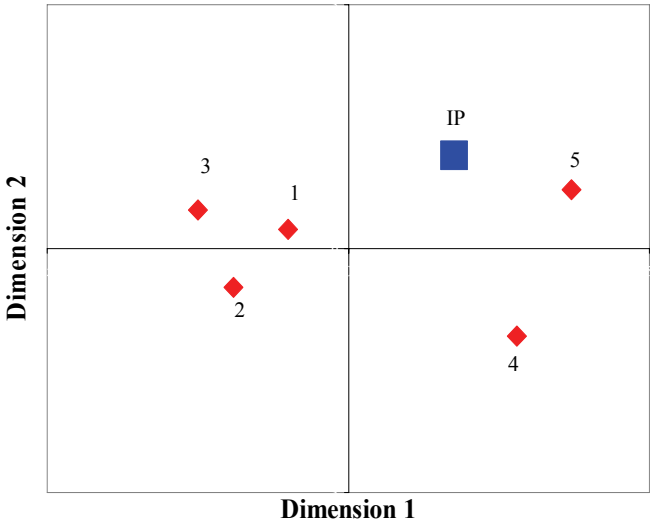


**Abbildung 3.15: MDS mit Einbettung eines Idealpunktes**

Die Einstellung einer Person zu einem Objekt ist durch die euklidische Distanz zum individuellen Idealpunkt gegeben (vgl. GIERL, 1995, S. 132). Damit ist jetzt eine Bewertung der Einstellung dieser Person möglich. Durch die Festlegung $\beta_3 = \beta_4$ bedeutet dies anschaulich, dass konzentrische Kreise um den Idealpunkt gelegt werden, wie Abbildung 3.16 verdeutlicht.

Als Nachteil der MDS bleibt die Gefahr einer Überforderung der Testpersonen bestehen. Sie wird durch die Tatsache verstärkt, dass für eine sinnvolle Berechnung der Modellparameter mindestens sechs, besser aber neun oder mehr Objekte beurteilt werden sollten (vgl. BACKHAUS et al., 2006, S. 646). (3.29) gibt den Datenverdichtungskoeffizienten Q in Abhängigkeit der Objekte (n) und Dimensionen (q) an, für den in der Literatur ein Wert größer 2 vorgeschlagen wird um sicherzustellen, dass die Parameter des Verfahrens sinnvoll geschätzt werden können (vgl. BACKHAUS et al., 2006, S. 646).

$$Q = \frac{n(n-1)/2}{nq} \geq 2 \qquad (3.29)$$

Bei Einhaltung des vorgeschlagenen Wertes des Datenverdichtungskoeffizienten sind für eine zweidimensionale Repräsentation bereits neun Objekte erforderlich. Dies entspricht 36 Paarvergleichen.

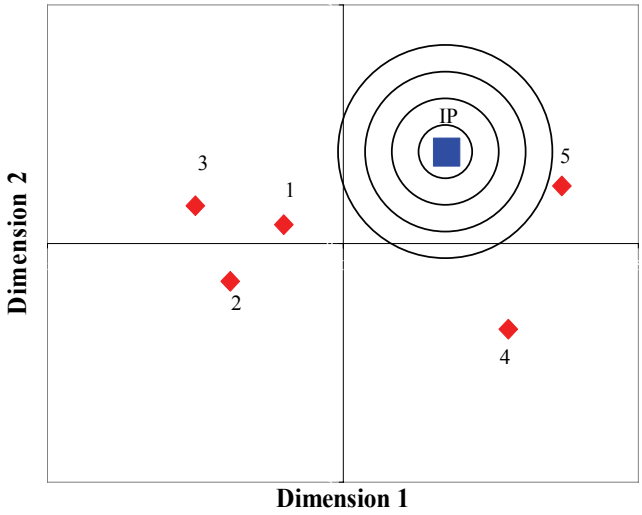


**Abbildung 3.16: Distanz der Objekte zum Idealpunkt**

### 3.3.3.5 Strukturgleichungsmodelle

Die bisher vorgestellten Skalierungsmethoden messen ein- oder mehrdimensionale Konstrukte. Strukturgleichungsmodelle versuchen dagegen gleichzeitig Zusammenhänge zwischen verschiedenen Konstrukten zu beschreiben (vgl. BACKHAUS et al., 2006, S. 338). Insofern gehen Strukturgleichungsmodelle über die reinen Skalierungsmethoden hinaus. Eine ausführliche Behandlung dieser Modelle ist jedoch angebracht, da die Messung latenter Variablen einer ihrer zentralen Bestandteile ist (vgl. HAIR et al., 2006, S. 712). Ausgangspunkt der Strukturgleichungsmodelle ist die *Kausalanalyse* zweier latenter Variablen (vgl. MEFFERT, 1992, S. 306). Diese Untersuchung der Abhängigkeit entspricht der Regressionsanalyse, allerdings mit der Einschränkung, dass die betrachteten Konstrukte nicht direkt messbar sind. Daher muss die Kausalanalyse um ein *Skalierungs- oder Messmodell* erweitert werden, welches mit Hilfe von beobachtbaren Variablen, in der Regel Items, die Konstrukte definiert (vgl.

BACKHAUS et al., 2006, S. 340). Abbildung 3.17 veranschaulicht diesen Zusammenhang für die latenten Variablen ´Einstellung´ und ´Kaufverhalten´.

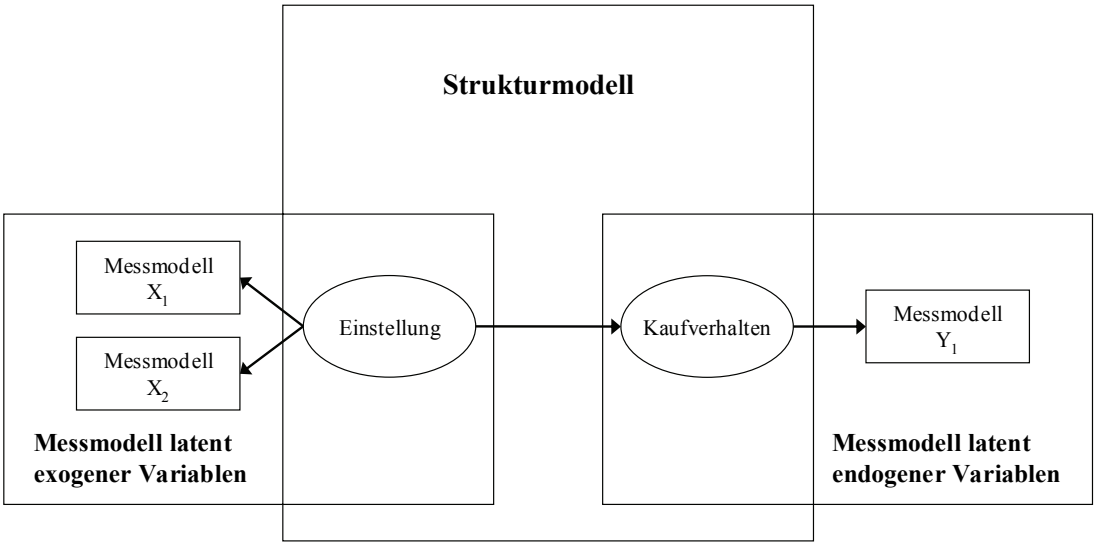


**Abbildung 3.17: Strukturgleichungsmodell bestehend aus Struktur- und Messmodell (in Anlehnung an BACKHAUS et al., 2006, S. 341)**

Die durch ein Strukturmodell abgebildeten Beziehungen zwischen beobachtbaren Variablen und Konstrukten bzw. zwischen zwei Konstrukten basieren auf sachlogischen oder theoretischen Hypothesen (vgl. BACKHAUS et al., 2006, S. 340). Damit entspringen Strukturgleichungsmodelle dem gleichen Denkmodell wie die konfirmatorische Faktorenanalyse, die im Prinzip einen Spezialfall der Strukturgleichungsmodelle darstellt (vgl. BACKHAUS et al., 2006, S. 330). Im Gegensatz zur explorativen Faktorenanalyse sind Strukturgleichungsmodelle den hypothesenprüfenden Verfahren zuzuordnen, das heißt sie dienen nicht der Entdeckung von Strukturen, sondern ihrer Überprüfung (vgl. BACKHAUS et al., 2006, S. 338). In der Literatur hat sich für die einzelnen Größen folgende Notation durchgesetzt (vgl. BACKHAUS et al., 2006, S. 348):

| Abkürzung | Bedeutung |
|---|---|
| $\eta$ | Latente endogene Variable, die im Modell erklärt wird |
| $\xi$ | Latente exogene Variable, die im Modell nicht erklärt wird |
| y | Manifeste Variable für eine latente endogene Variable |
| x | Manifeste Variable für eine latente exogene Variable |
| $\varepsilon$ | Residualvariable für eine manifeste Variable y |
| $\delta$ | Residualvariable für eine manifeste Variable x |
| $\zeta$ | Residualvariable für eine latente endogene Variable $\eta$ |

**Tabelle 3.5: Abkürzungen der Variablen in einem Strukturgleichungsmodell (in Anlehnung an BACKHAUS et al., 2006, S. 349)**

Zwischen zwei beobachtbaren Variablen $X_1$ und $X_2$ sind mehrere Abhängigkeitsbeziehungen vorstellbar. Wenn $X_1$ und $X_2$ beispielsweise zwei Items desselben Konstrukts sind, kann neben dem tatsächlichen wechselseitigen Einfluss auch die latente Variable die Ursache für die Abhängigkeit sein (vgl. BACKHAUS et al., 2006, S. 346). In diesem Fall spricht man von einer Scheinkorrelation zwischen $X_1$ und $X_2$ (vgl. BAMBERG, BAUR, 2002, S. 50). Abbildung 3.18 zeigt exemplarisch einen möglichen Wirkungszusammenhang zwischen zwei Items $X_1$ und $X_2$ und einem Konstrukt $\xi$.

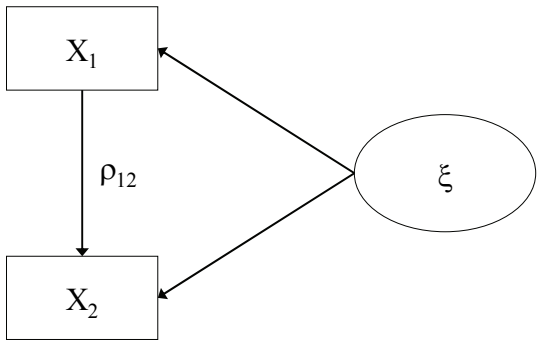


**Abbildung 3.18: Mögliche Abhängigkeit zwischen latenter und manifesten Variablen (in Anlehnung an BACKHAUS et al., 2006, S. 346)**

Die Korrelation $\rho_{12}$ entspricht nur zum Teil dem kausalen Zusammenhang, da $X_2$ nicht nur direkt von $X_1$ beeinflusst wird, sondern vom Konstrukt $\xi$ sowohl direkt, als auch indirekt über $X_1$ (vgl. BACKHAUS et al., 2006, S. 346). Um den tatsächlichen Zusammenhang zwischen $X_1$ und $X_2$ zu bestimmen, muss $\xi$ konstant gehalten werden, was durch die Berechnung des partiellen Korrelationskoeffizienten möglich ist (HARTUNG, ELPELT, 1992, S. 181):

$$\rho_{(X_1X_2)|\xi} = \frac{\rho_{X_1X_2} - \rho_{X_1\xi}\rho_{X_2\xi}}{\sqrt{(1-\rho^2_{X_1\xi})(1-\rho^2_{X_2\xi})}} \tag{3.30}$$

Die Unterscheidung zwischen direktem und indirektem Einfluss ist ein zentraler Aspekt der Strukturgleichungsmodelle. Allerdings sind dabei Kausalbeziehungen zwischen Konstrukten und nicht zwischen Items relevant. So kann analog zu Abbildung 3.18 auch der Wirkungszusammenhang zwischen drei Konstrukten, wie in Abbildung 3.19 dargestellt werden:

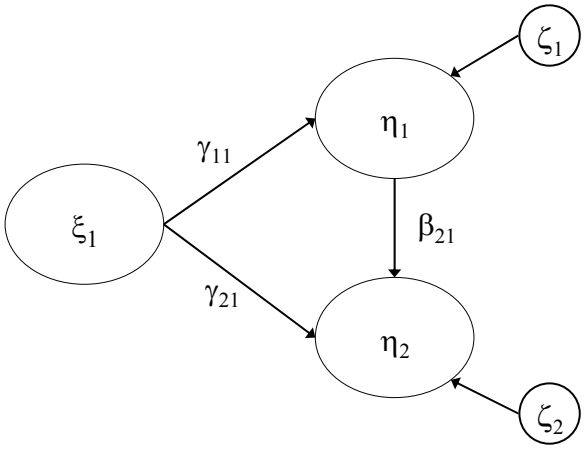


**Abbildung 3.19: Beispiel für ein Strukturmodell (in Anlehnung an BACKHAUS et al., 2006, S. 349)**

$\zeta$ ist die Residualvariable der jeweiligen endogenen latenten Variable. Sind alle Größen standardisiert, dann können die endogenen Konstrukte $\eta_1$ und $\eta_2$ mittels einer Regressionsgleichung ohne Konstante wie in (3.31) beschrieben werden. Die Konstante kann entfallen, da standardisierte Merkmale einen Erwartungswert von Null aufweisen. Die Matrixschreibweise erklärt auch die nicht zwangsläufig intuitive Indizierung von $\beta$ und $\gamma$ (vgl. BACKHAUS et al., 2006, S. 350).

$$\eta_1 = \gamma_{11}\xi_1 + \zeta_1$$
$$\eta_2 = \beta_{21}\eta_1 + \gamma_{21}\xi_1 + \zeta_2$$

$$\begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix} = \begin{pmatrix}0 & 0\\ \beta_{21} & 0\end{pmatrix}\begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix} + \begin{pmatrix}\gamma_{11}\\ \gamma_{21}\end{pmatrix}\xi_1 + \begin{pmatrix}\zeta_1\\ \zeta_2\end{pmatrix} \tag{3.31}$$

$$\boldsymbol{\eta} = \boldsymbol{B\eta} + \boldsymbol{\Gamma\xi} + \boldsymbol{\zeta}$$

Die Standardisierung hat den zusätzlichen Vorteil, dass die Höhe der Regressionskoeffizienten Aussagekraft über die Stärke des Einflusses besitzt. Allerdings sind die in (3.31) dargestellten Zusammenhänge empirisch nicht überprüfbar. Deswegen müssen für das endogene Konstrukt $\xi_1$ und die exogenen Konstrukte $\eta_1$ und $\eta_2$ Messmodelle unterstellt werden (vgl. BACKHAUS et al., 2006, S. 355). Aber auch das in Abbildung 3.19 gezeigte Abhängigkeitsverhältnis zwischen den Konstrukten $\xi_1$, $\eta_1$ und $\eta_2$ beruht lediglich auf sachlogisch begründeten Annahmen (vgl. BACKHAUS et al., 2006, S. 355). Dies unterstreicht den konfirmatorischen Charakter der Strukturgleichungsmodelle, da die vermuteten Zusammenhänge zwischen Items und Konstrukten sowie zwischen zwei Konstrukten festgelegt sind und durch das Modell lediglich überprüft werden (vgl. BACKHAUS et al., 2006, S. 356). Für $\xi_1$ könnte beispielsweise das Messmodell in Abbildung 3.20 unterstellt werden:

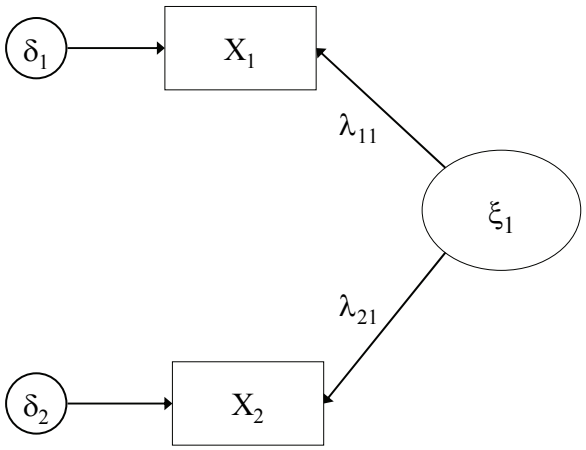


**Abbildung 3.20: Beispiel für ein Messmodell (in Anlehnung an BACKHAUS et al., 2006, S. 350)**

Dabei ist zu beachten, dass die latente Variable $\xi_1$ Einfluss auf die beobachtbaren Variablen nimmt und nicht umgekehrt. Dies entspricht exakt der Denkweise der

Faktorenanalyse, weshalb Strukturgleichungsmodelle eine Mischung aus Regressions- und Faktorenanalyse darstellen (vgl. BACKHAUS et al., 2006, S. 351). Wenn alle Variablen standardisiert sind, kann das Messmodell für $\xi_1$ auch durch eine Regressionsgleichung abgebildet werden (vgl. BACKHAUS et al., 2006, S. 350).

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} \lambda_{11} \\ \lambda_{21} \end{pmatrix} \xi_1 + \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix} \qquad (3.32)$$

$$\boldsymbol{X} = \boldsymbol{\Lambda_X \xi} + \boldsymbol{\delta}$$

Analog läuft die Berechnung für die endogenen Variablen:

$$\boldsymbol{Y} = \boldsymbol{\Lambda_Y \eta} + \boldsymbol{\varepsilon} \qquad (3.33)$$

Nach Aufstellung eines Messmodells für alle Konstrukte und nach Beschreibung aller vermuteten Zusammenhänge zwischen zwei Konstrukten, gelangt man beispielsweise zum in Abbildung 3.21 dargestellten vollständigen Pfaddiagramm.

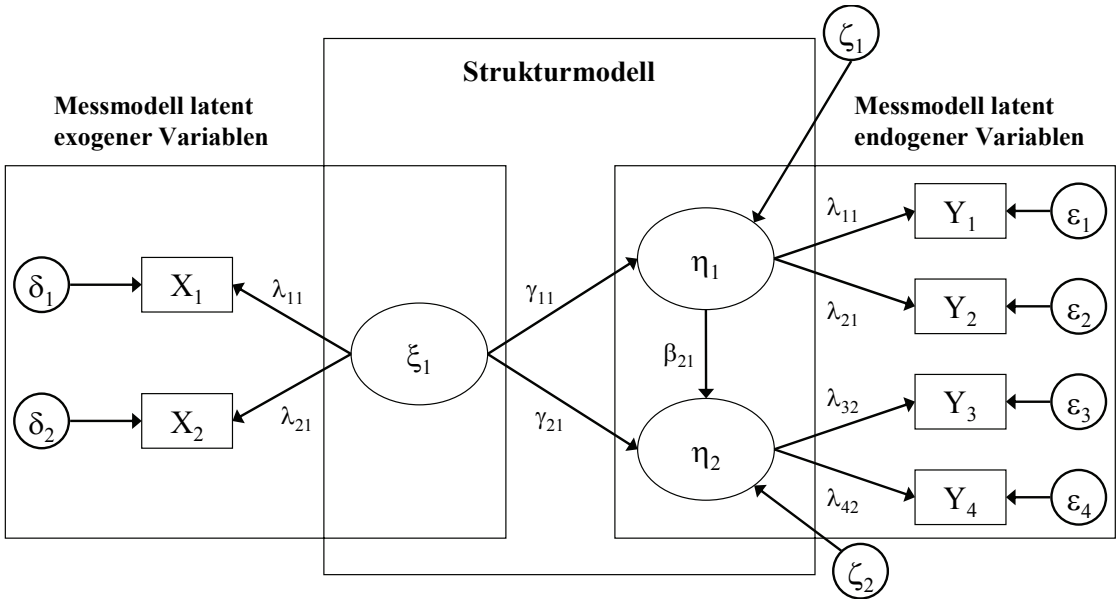


**Abbildung 3.21: Pfadmodell eines vollständigen Strukturgleichungsmodells (in Anlehnung an BACKHAUS et al., 2006, S. 355)**

In Abbildung 3.21 sind sowohl die Verbindungen zwischen $\mathbf{Y}$ und $\boldsymbol{\eta}$, als auch die Verbindungen zwischen $\mathbf{X}$ und $\boldsymbol{\xi}$ mit $\lambda$ bezeichnet. Dies bedeutet jedoch nicht, dass diese Parameter den gleichen Wert annehmen, da zwischen $\boldsymbol{\Lambda}_\mathbf{X}$ und

$\boldsymbol{\Lambda}_{\mathbf{Y}}$ unterschieden wird. Die Schätzung der Parameter erfolgt in Abhängigkeit der gewählten Zusammenhänge analog zu (3.31), (3.32) und (3.33).

Die Parameter können als *freie* Parameter aus den empirischen Daten geschätzt werden (vgl. BACKHAUS et al., 2006, S. 365). Bei *restringierten* Parametern ist auf Grund theoretischer Überlegungen zusätzliche Information vorhanden, beispielsweise dass zwei Parameter gleich groß sind (vgl. BACKHAUS et al., 2006, S. 365). Bei *festen* Parametern ist der Einfluss exakt bekannt, etwa wenn kausale Beziehungen zwischen zwei Größen ausgeschlossen werden und der Parameter daher gleich Null gesetzt wird (vgl. BACKHAUS et al., 2006, S. 365). Restringierte und feste Parameter reduzieren die Zahl der zu schätzenden Parameter. Als Schätzverfahren kommen die Maximum-Likelihood-Methode, verschiedene Kleinst-Quadrate-Schätzer und asymptotisch verteilungsfreie Schätzer in Betracht (vgl. BACKHAUS et al., 2006, S. 368), auf die an dieser Stelle jedoch nicht näher eingegangen wird.

## 3.4 Qualität der Messergebnisse

Die Qualität der Messergebnisse ist ein Hinweis dafür, ob Messinstrumente die zu messende Größe adäquat messen. Die zu messenden Größen können insbesondere Konstrukte oder Dimensionen von Konstrukten sein. Ausgangspunkt der Beurteilung der Qualität der Messergebnisse ist folgende Überlegung: Bei einer perfekten Messung entspricht das Gemessene exakt der tatsächlichen Größe, z.B. der Einstellung der Person zu einem Produkt. Je weiter die gemessene Größe von der tatsächlichen Größe abweicht, desto schlechter ist die Messung (vgl. SCHNELL et al., 2005, S. 149). Formal ergibt sich damit:

$$X_O = X_T + E \tag{3.34}$$

$X_O$: *Gemessener Wert*

$X_T$: *Wahrer Wert*

$E$: *Messfehler*

Die Bestimmung der Qualität der Messergebnisse wäre sehr einfach, wenn der wahre Wert $X_T$ bekannt wäre. Dieser Fall ist in der Praxis jedoch so gut wie

ausgeschlossen. Daher muss die Bestimmung der Qualität auf Umwegen erfolgen.

- Dazu sind zunächst methodische Grundlagen erforderlich, auf die bei der Beurteilung der Qualität zurückgegriffen wird (Kapitel 3.4.1).

Mit Hilfe dieser methodischen Grundlagen können dann die *Reliabilität* und *Validität* der Messung bestimmt werden.

- Die Reliabilität beschreibt die Zuverlässigkeit einer Messung. Die Reliabilität einer Messung ist hoch, wenn mehrere Messungen zu ähnlichen Ergebnissen kommen. Im Idealfall liefern alle Messungen dasselbe Ergebnis. Das heißt, die Messung sollte frei von Zufallsfehlern sein (vgl. SCHNELL et al., 2005, S. 151). In Kapitel 3.4.2 wird ausführlich auf die Bestimmung der Reliabilität eingegangen.

- Die Validität gibt die Gültigkeit einer Messung an. Hier steht demnach die Frage im Mittelpunkt, ob der Sachverhalt gemessen wird, der gemessen werden soll. Die Validität der Messung ist hoch, wenn nur ein geringer systematischer Fehler auftritt. Im Idealfall existiert kein systematischer Fehler (siehe Kapitel 3.4.3).

- Neben der Reliabilität und Validität existieren weitere Kriterien zur Beurteilung der Qualität der Messergebnisse. Diese sind in Kapitel 3.4.4 erläutert.

Die beiden zentralen Maße zur Beurteilung der Messergebnisse sind Reliabilität und Validität. Um diese eindeutig voneinander abzugrenzen, ist eine Präzisierung von (3.34) erforderlich:

$$X_O = X_T + X_S + X_R \tag{3.35}$$

$X_O$: *Gemessener Wert*

$X_T$: *Wahrer Wert*

$X_S$: *Systematischer Fehler*

$X_R$: *Zufälliger Fehler*

Eine reliable Messung ist gleichbedeutend mit $X_R = 0$ (Bild 1 und 2 in Abbildung 3.22). Die Messung ist valide, wenn zusätzlich $X_S = 0$ ist (vgl. SCHNELL et al.,

2005, S. 154). Eine Messung ist demzufolge nur dann valide, wenn $X_R = 0$ und $X_S = 0$ ist (siehe Bild 1 in Abbildung 3.22):

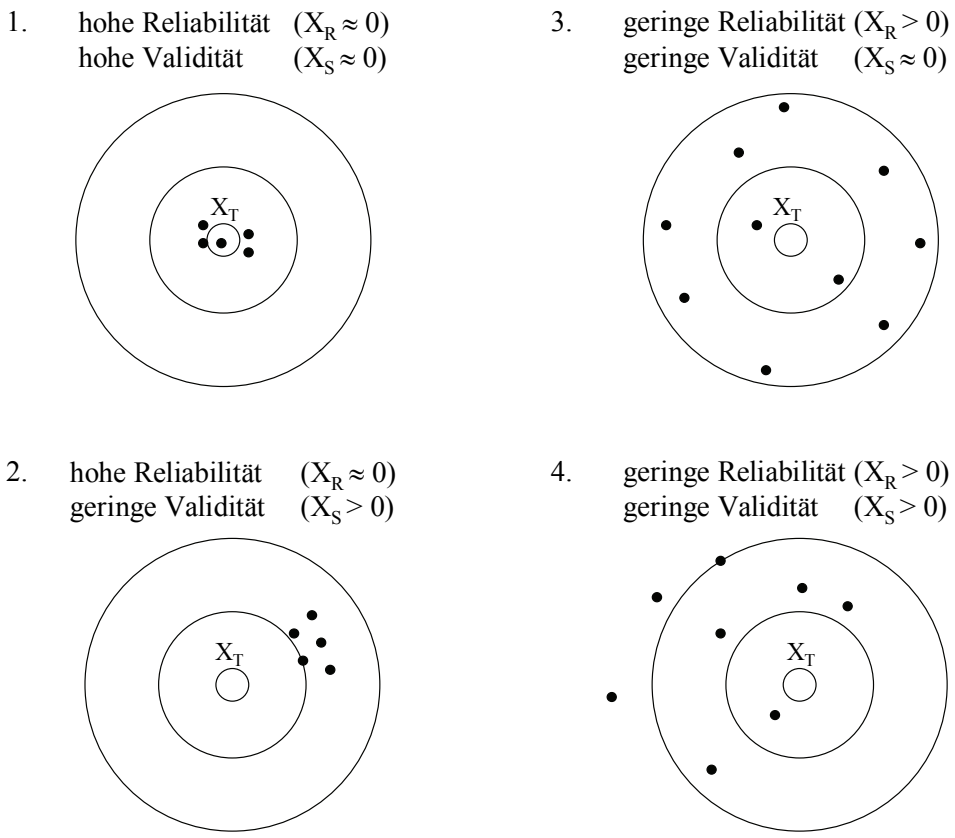


**Abbildung 3.22: Auswirkung von Reliabilität und Validität auf Messergebnisse (vgl. TROCHIM, 2007)**

Ist eine Messung nicht reliabel, aber dennoch frei von einem systematischen Fehler $X_S$ (Bild 3), verbessern Messwiederholungen oder die Verwendung mehrerer ähnlicher Items die Messergebnisse. Der Mittelwert dieser Messungen liegt dann nahe am wahren Wert $X_T$, da der Zufallsfehler keine Verschiebung in eine bestimmte Richtung verursacht, weshalb die Abweichungen sich gegenseitig aufheben. Falls jedoch zusätzlich $X_S$ ungleich Null ist (Bild 4), führt die mehrmalige Wiederholung der Messung zu keiner Verbesserung der Ergebnisse.

## 3.4.1 Methodische Grundlagen

Viele der in diesem Kapitel vorgestellten Maße zur Beurteilung von Messergebnissen nutzen den Zusammenhang zweier Größen als Grundlage. Neben dem mit Hilfe von Korrelationsmaßen gemessenen Zusammenhang zweier Merkmale kann auch eine graphische Veranschaulichung der Messergebnisse auf Basis von Boxplots hilfreich sein, wie sie in Kapitel 3.4.4.3 näher beschrieben wird. In diesem Kapitel sollen deshalb zunächst verschiedene Korrelationsmaße (Kapitel 3.4.1.1) und Boxplots (Kapitel 3.4.1.2) allgemein erklärt werden.

### 3.4.1.1 Korrelationsmaße

Der Zusammenhang zweier Größen kann mit Hilfe ihrer Korrelation bestimmt werden. Bei kardinal bzw. ordinal skalierten Merkmalen sind folgende Korrelationsmaße üblich:

- Bravais-Pearson Korrelationskoeffizient
- Rangkorrelationskoeffizient von Spearman
- Kendalls $\tau$

Alle Korrelationsmaße können nicht nur die Stärke des Zusammenhangs, sondern zusätzlich die Richtung bestimmen (vgl. BAMBERG, BAUR, 2002, S. 36). Sie bewegen sich in einem Intervall von -1 bis +1. Bei einem sehr starken positiven Zusammenhang erreichen die Koeffizienten nahezu den Wert +1, bei einer gegenläufigen Beziehung dagegen näherungsweise -1. Ein Wert nahe 0 deutet darauf hin, dass zwei Merkmale unkorreliert sind.

Das bekannteste Maß zur Bestimmung der Abhängigkeit zweier Merkmale ist der *Bravais-Pearson-Korrelationskoeffizient*, der allerdings nur für metrische Daten zulässig ist.

**Definition 3.3**

*Für zwei kardinal skalierte Merkmale x und y, die bei n Objekten gemessen werden, heißt der Ausdruck*

$$r^{BP} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i ,$$

***Bravais-Pearson-Korrelationskoeffizient*** *(vgl. BAMBERG, BAUR, 2002, S. 36).*

Neben diesem Korrelationsmaß für metrische Daten kann auch auf Zusammenhangsmaße für ordinale Daten zurückgegriffen werden. Das bekannteste Maß ist der Rangkorrelationskoeffizient von Spearman. Bei ihm werden allen Objekten auf Basis ihrer Merkmalsausprägungen Rangplätze zugewiesen. Diese Rangplätze werden als metrische Ausprägung betrachtet und zur Berechnung des bekannten Bravais-Pearson-Korrelationskoeffizienten herangezogen:

**Definition 3.4**

*Für zwei mindestens ordinale Merkmale X und Y ist für den Fall ohne Bindungen der* ***Rangkorrelationskoeffizient von Spearman*** *wie folgt definiert (vgl. BAMBERG, BAUR, 2002, S. 38):*

$$r^{SP} = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

$d_i = Rg(x_i) - Rg(y_i)$ *für* $i = 1, \ldots, n.$

Insbesondere bei ratingbasierten Daten kommt es häufig zu identischen Ausprägungen, so genannten Bindungen, deren Existenz eine Abwandlung der geläufigen Berechnung des Rangkorrelationskoeffizienten erforderlich macht (*vgl. HARTUNG, ELPELT, 1992, S. 192).* Neben dem Rangkorrelationskoeffizienten existiert mit Kendalls τ noch ein weiteres ordinales Korrelationsmaß. Prinzipiell beruht Kendalls τ auf der Idee des paarweisen Vergleichs der Objekte. Dazu werden alle konkordanten und diskordanten Objektpaare gezählt.

**Definition 3.5**

*Gegeben seien zwei Merkmale X und Y. Jedes Objektpaar (i,j), welches folgende Eigenschaft erfüllt, heißt* ***konkordant****:*

$$x_i < x_j \Rightarrow y_i < y_j \qquad \textit{bzw.} \qquad x_i > x_j \Rightarrow y_i > y_j$$

*Ein Objektpaar heißt dagegen* ***diskordant****, wenn gilt:*

$$x_i < x_j \Rightarrow y_i > y_j \qquad \textit{bzw.} \qquad x_i > x_j \Rightarrow y_i < y_j$$

In Abhängigkeit des Untersuchungsdesigns existieren drei Formen von Kendalls τ. Da Kendalls τ-a für Merkmale mit Bindungen ungeeignet ist und deshalb nur in Ausnahmefällen eingesetzt werden kann, soll es hier nicht weiter betrachtet werden (vgl. KENDALL, 1970, S. 34-36). Ein geeignetes Maß beim Vorhandensein von Bindungen ist Kendalls τ-b. Zur Berechnung von Kendalls τ-b müssen zunächst alle konkordanten und diskordanten Objektpaare summiert werden. Zusätzlich müssen bei beiden Merkmalen die Objektpaare mit gleicher Merkmalsausprägung gezählt werden.

**Definition 3.6**

*Der Ausdruck*

$$r^{\tau-b} = \frac{P-Q}{\sqrt{(\frac{1}{2}n\cdot(n-1)-X_T)}\cdot\sqrt{(\frac{1}{2}n\cdot(n-1)-Y_T)}},$$

*heißt* ***Kendalls τ-b****. Dabei ist*

*P:* *Anzahl der konkordanten Paare*

*Q:* *Anzahl der diskordanten Paare*

*n:* *Anzahl der Objekte*

$X_T$: *Anzahl der Bindungen bei Merkmal X*

$Y_T$: *Anzahl der Bindungen bei Merkmal Y.*

Allerdings besitzt auch Kendalls τ-b Eigenschaften, die in manchen Fällen unerwünscht sind. Beispielsweise kann Kendalls τ-b nicht die Werte -1 und +1 annehmen, falls die beiden Merkmale unterschiedlich viele Ausprägungen aufweisen oder die Anzahl der Objekte n kein Vielfaches der

Merkmalsausprägungen sind (vgl. KENDALL, 1970. S. 47). Daher führt KENDALL eine weitere Maßzahl ein, Kendalls τ-c.

**Definition 3.7**

$$r^{\tau-c} = \frac{2 \cdot (P - Q)}{n^2 \cdot \frac{min\{k,l\} - 1}{min\{k,l\}}},$$

*P:* *Anzahl der konkordanten Paare*

*Q:* *Anzahl der diskordanten Paare*

*n:* *Anzahl der Objekte*

*k:* *Anzahl der Ausprägungen bei Merkmal X (k > 1)*

*l:* *Anzahl der Ausprägungen bei Merkmal Y (l > 1)*

Kendalls τ-c kann auch für $k \neq l$ die Werte ±1 nahezu erreichen. Falls n jedoch kein Vielfaches vom min (i,j) ist, nimmt auch diese Kennzahl nicht die Extremwerte ±1 an (vgl. HILBERT, 1998, S. 141).

### 3.4.1.2 Boxplots

Boxplots sind gängige graphische Verfahren zur komprimierten Darstellung der Verteilung eines Merkmals (vgl. HARTUNG, ELPELT, 1992; S. 597ff.). Dazu müssen der Median und verschiedene Quantile (Perzentile) der Verteilung bekannt sein. Ein p-Quantil $x_p$ zerlegt die Merkmalswerte dabei so, dass mindestens ein Anteil p der Werte kleiner oder gleich $x_p$ ist und ein Anteil (1-p) der Werte größer oder gleich $x_p$ ist. Das bekannteste Quantil ist der Median ($x_{0,5}$), der dem mittleren Merkmalswert entspricht. Das heißt, 50% der Merkmalswerte sind kleiner oder gleich und 50% der Merkmalswerte sind größer oder gleich dem Median.

Mit Hilfe der Quantile kann die Form eines Boxplots bestimmt werden. Die innere Box ist durch das erste und dritte Quartil ($x_{0,25}$ und $x_{0,75}$) begrenzt, den Median kennzeichnet ein senkrechter Strich. Darüber hinaus können Ausreißer speziell gekennzeichnet werden. In Abbildung 3.23 ist ein Boxplot exemplarisch dargestellt:

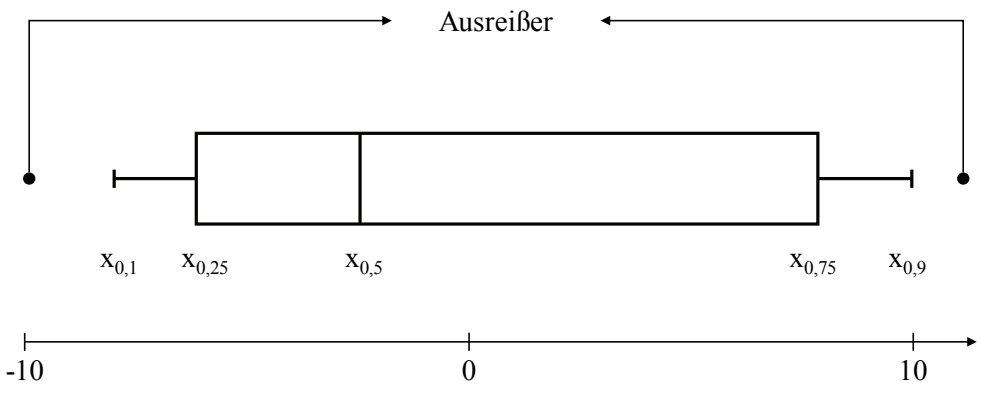


**Abbildung 3.23: Darstellung der Merkmalswerte mit Hilfe eines Boxplots**

An Hand des Boxplots in Abbildung 3.23 ist ersichtlich, dass für dieses Merkmal der Wert von $x_{0,9}$ gleich 10 ist. Der Median liegt bei ca. -2,5. Entsprechend können auch die anderen Quantilswerte dieser Verteilung interpretiert werden.

## 3.4.2 Messung der Reliabilität

In Abbildung 3.22 (Seite 76) wurde bereits gezeigt, dass eine Messung selbst dann reliabel sein kann, wenn systematische Fehler auftreten. Daher kann bei den Überlegungen zur Prüfung der Reliabilität der systematische Fehler unberücksichtigt bleiben. Zur Überprüfung des Fehlens eines Zufallsfehlers ist es zunächst erforderlich, diesen genau zu definieren. In der klassischen Testtheorie geht man von folgenden Annahmen für $X_R$ aus (vgl. SCHNELL et al., 2005, S. 150):

- Der Mittelwert des Zufallsfehlers ist Null ($\overline{X}_R = 0$). Deshalb ist durch Mittelung mehrerer Messergebnisse eine Verbesserung erzielbar, wenn kein systematischer Fehler vorliegt.

- Die Korrelation zwischen zufälligem Messfehler und wahrem Messwert ist Null ($r_{X_R X_T} = 0$). Dies bedeutet, dass die Höhe des Messfehlers nicht von der Höhe der Ausprägung abhängen darf.

- Die zufälligen Messfehler verschiedener Messungen (i,j) sind unkorreliert ($r_{X_{Ri} X_{Rj}} = 0$). Diese Forderung hat weit reichende Konsequenzen, da somit

die zufälligen Messfehler zweier unterschiedlicher Items unkorreliert sein müssen.

- Der zufällige Messfehler korreliert nicht mit dem wahren Wert einer anderen Messung ($r_{X_{Ri}X_{Tj}} = 0$). Es sind demzufolge keine systematischen Einflüsse anderer Variablen möglich.

Diese Voraussetzungen bedeuten jedoch nicht, dass die beobachteten Werte konstant sein müssen. Falls der wahre Wert Schwankungen unterworfen ist, muss der beobachtete Wert ebenfalls variieren. Die reliable Messung des Pegelstandes eines Flusses schwankt synchron zu den Wasserständen. Die Reliabilität ist somit das Verhältnis von wahrer zu beobachteter Varianz (vgl. SCHNELL et al., 2005, S. 151). Da die Fehlerterme und die wahren Ausprägungen als unkorreliert vorausgesetzt werden, ist die beobachtete Varianz gleich der Summe aus wahrer Varianz und Fehlervarianz. Für die Reliabilität kann damit folgendes festgehalten werden:

$$R = \frac{s^2_{X_T}}{s^2_{X_O}} = \frac{s^2_{X_T}}{s^2_{X_T} + s^2_{X_S} + s^2_{X_R}} \tag{3.36}$$

$s^2_{XY}$: *Wahre Varianz*

$s^2_{XO}$: *Beobachtete Varianz*

$s^2_{XS}$: *Varianz des systematischen Fehlers*

$s^2_{XR}$: *Varianz des zufälligen Fehlers*

Die Reliabilität nimmt für $s^2_{X_R} = 0$ den Wert Eins an, wenn der systematische Fehler einer Messung konstant und damit $s^2_{X_S} = 0$ ist. Die Reliabilität sinkt, wenn die Fehlervarianz im Vergleich zu Varianz des wahren Wertes größer wird. (3.36) verdeutlicht zwar, wie ein Reliabilitätskoeffizient definiert ist, zur Berechnung ist (3.36) jedoch ungeeignet, da die Varianzanteile unbekannt sind (vgl. SCHNELL et al., 2005, S. 151). Daher greift man zur Messung der Reliabilität auf die Test-Retest-Methode und die Paralleltestmethode zurück.

Die *Test-Retest-Methode* misst die zeitliche Stabilität der Ergebnisse. Dazu muss dasselbe Item zu zwei verschiedenen Zeitpunkten $t_1$ und $t_2$ denselben Personen vorgelegt werden, wie Abbildung 3.24 veranschaulicht:

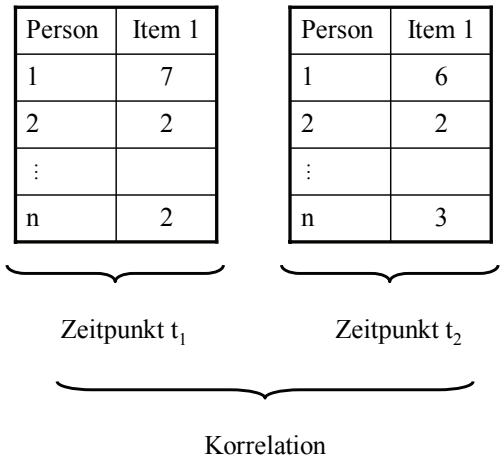


**Abbildung 3.24: Vorgehensweise bei der Test-Retest-Methode**

Die Korrelation der beiden Messungen liefert ein Maß für die Reliabilität (vgl. SCHNELL et al., 2005, S. 151).

$$R = r_{X_{t1} X_{t2}} \tag{3.37}$$

Diese Methode unterstellt, dass $X_T$ zwischen den beiden Messungen konstant bleibt. Daher unterschätzt sie die tatsächliche Reliabilität, wenn $X_T$ zwischen $t_1$ und $t_2$ Schwankungen unterworfen ist. Bei der Test-Retest-Methode besteht jedoch auch die Gefahr, die Reliabilität einer Messung zu überschätzen. Dies ist der Fall, wenn Versuchspersonen beim Retest noch wissen, wie sie bei der ersten Befragung geantwortet haben. Wegen der Gefahren der Über- bzw. Unterschätzung der tatsächlichen Reliabilität hat sich diese Methode in der Praxis nicht durchgesetzt (vgl. SCHNELL et al., 2005, S. 152).

Die *Paralleltestmethode* unterstellt, dass mehrere Messinstrumente (z.B. Items) gleichermaßen die zu messende Größe beschreiben können. Wenn zwei Items exakt das Gleiche messen, dann liefert die Korrelation dieser Items ein Maß für die Reliabilität (vgl. SCHNELL et al., 2005, S. 152). In der Praxis werden dabei häufig wesentlich mehr als zwei Items eingesetzt. Bei mehr als zwei Items werden zur Berechnung der Reliabilität die Items zufällig in zwei Hälften ($K_1$ und $K_2$) geteilt (*Splithalf-Methode*). Anschließend wird für jede Person ein Gesamtscore pro Itemgruppe durch Mittelwertbildung berechnet (vgl. JANSSEN, LAATZ, 2005, S. 566). Die Korrelation dieser Gesamtscorewerte entspricht der Reliabilität der Messung:

$$R_{K_1} = R_{K_2} = R = r_{K_1 K_2} \tag{3.38}$$

In Abbildung 3.25 ist die Vorgehensweise bei der Splithalf-Methode dargestellt. Man geht davon aus, dass die dort dargestellten Items alle dasselbe Konstrukt messen. Anschließend werden die Items in zwei Gruppen zufällig eingeteilt, z.B. Item 1-3 zur Itemgruppe 1 und Item 4-6 zur Itemgruppe 2. Mit Hilfe dieser Itemgruppen wird anschließend ein Maß für die Reliabilität bestimmt.

| Person | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 | Item 6 |
|---|---|---|---|---|---|---|
| 1 | 7 | 6 | 7 | 5 | 6 | 4 |
| 2 | 2 | 1 | 1 | 2 | 1 | 1 |
| ⋮ | | | | | | |
| n | 2 | 4 | 5 | 5 | 4 | 5 |

Itemgruppe 1 Itemgruppe 2

⇩ ⇩

| Person | $K_1$ | $K_2$ |
|---|---|---|
| 1 | 6,7 | 5,0 |
| 2 | 1,3 | 1,3 |
| ⋮ | | |
| n | 3,7 | 4,7 |

Korrelation

**Abbildung 3.25: Vorgehensweise bei der Splithalf-Methode**

Das Ergebnis für die Reliabilität hängt bei der Splithalf-Methode von der zufälligen Aufteilung der Items ab (vgl. SCHNELL et al., 2005, S. 153). Diesen Nachteil behebt der Koeffizient *Cronbach´s α*, da hier der Mittelwert aller möglichen Itemaufteilungen als Maß für die Reliabilität verwendet wird. Eine vereinfachte Darstellung für Cronbach´s α ist unter der Voraussetzung möglich, dass alle Items identische Varianzen haben. In diesem Fall hängt Cronbach´s α ausschließlich von der Anzahl und der durchschnittlichen Korrelation der Items ab. Die Annahme gleicher Varianzen scheint insbesondere bei Ratingskalen gleicher Skalenbreite gerechtfertigt, weshalb an dieser Stelle lediglich diese vereinfachte Form von Cronbach´s α aufgeführt werden soll (vgl. SCHNELL et al., 2005, S. 153):

$$\alpha = \frac{m\bar{r}_{ij}}{1+(m-1)\bar{r}_{ij}} \tag{3.39}$$

*m:* *Anzahl der Items*

$\bar{r}_{ij}$ *:* *Mittelwert aller Korrelationen der Itempaare (i,j)*

In Abbildung 3.26 ist das Vorgehen bei der Bestimmung des Mittelwerts aller Korrelationen schematisch dargestellt:

| Person | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 | Item 6 |
|---|---|---|---|---|---|---|
| 1 | 7 | 6 | 7 | 5 | 6 | 4 |
| 2 | 2 | 1 | 1 | 2 | 1 | 1 |
| ⋮ | | | | | | |
| n | 2 | 4 | 5 | 5 | 4 | 5 |

**Korrelationsmatrix**

| $r_{ij}$ | Item 1 | Item 2 | Item 3 | Item 4 | Item 5 | Item 6 |
|---|---|---|---|---|---|---|
| Item 1 | | 0,8 | 0,69 | 0,77 | 0,65 | 0,9 |
| Item 2 | | | 0,58 | 0,69 | 0,58 | 0,88 |
| Item 3 | | | | 0,66 | 0,92 | 0,81 |
| Item 4 | | | | | 0,67 | 0,58 |
| Item 5 | | | | | | 0,7 |
| Item 6 | | | | | | |

$\bar{r}_{ij}$

**Abbildung 3.26: Vorgehensweise bei der Berechnung von Cronbach´s α**

Aus Formel (3.39) wird ersichtlich, dass Cronbach´s α steigt, wenn zusätzlich berücksichtigte Items die durchschnittliche Korrelation nicht verändern. Daher nimmt Cronbach´s α selbst bei sehr kleinen durchschnittlichen Korrelationen sehr große Werte an, wenn viele Items einbezogen werden (vgl. SCHNELL et al., 2005, S. 153). Diese Eigenschaft von Cronbach´s α ist für die Messung der Reliabilität sinnvoll, da davon ausgegangen werden kann, dass der zufällige Fehler einer Messung abnimmt, wenn mehrere Items verwendet werden (vgl. Abbildung 3.22, Seite 76). Allerdings lässt Cronbach´s α keine Aussage über einen systematischen Messfehler zu. Dieser kann nur mit Hilfe der Methoden zur Messung der Validität bestimmt werden.

### 3.4.3 Messung der Validität

Mit Hilfe der Reliabilität wurde überprüft, ob ein Messergebnis frei von zufälligen Fehlern ist. Die Validität beschreibt nun das Ausmaß, mit dem ein Messinstrument tatsächlich misst, was es messen soll (vgl. SCHNELL et al., 2005, S. 154). Die Abgrenzung zur Reliabilität veranschaulicht das Beispiel eines falsch geeichten Meterstabes. Dieser wird zwar stets identische Ergebnisse für gleich große Personen liefern und somit perfekt reliabel sein, dennoch misst der Meterstab nicht den zu messenden Sachverhalt. Daher sind die Ergebnisse nicht valide.

Für die Bewertung der Validität stehen drei Ansatzpunkte zur Verfügung (vgl. HAMMANN, ERICHSON, 2000, S. 77), die Inhaltsvalidität, die Kriteriumsvalidität und die Konstruktvalidität.

Die *Inhaltsvalidität* besagt, dass alle Aspekte des zu messenden Konstrukts oder der zu messenden Dimension berücksichtigt werden müssen (vgl. SCHNELL et al., 2005, S. 155). Diese Aspekte sind lediglich subjektiv erfassbar, weshalb die Inhaltsvalidität kein objektives Kriterium darstellt (vgl. HAMMANN, ERICHSON, 2000, S. 77). Zum Teil wird in der Literatur daher die Auffassung vertreten, es handle sich hierbei lediglich um eine bei der Konzeptualisierung zu berücksichtigende Forderung und nicht um ein Bewertungskriterium (vgl. SCHNELL et al., 2005, S. 155).

Die *Kriteriumsvalidität* (Kapitel 3.4.3.1) und die *Konstruktvalidität* (Kapitel 3.4.3.2) sind dagegen allgemein akzeptierte Bewertungskriterien. Daher soll auf diese beiden im Folgenden explizit eingegangen werden.

#### 3.4.3.1 Kriteriumsvalidität

Die Kriteriumsvalidität bezieht sich auf die Beziehung zwischen dem Messinstrument und einem externen Kriterium (vgl. HAMMANN, ERICHSON, 2000, S. 77). Das externe Kriterium kann ein einzelner Indikator (Item, Paarvergleich) oder das, auf andere Weise gemessene, selbe Konstrukt sein. Die Kriteriumsvalidität liefert nur dann sinnvolle Aussagen über die Qualität eines Messinstruments, wenn bei der Messung des externen Kriteriums keine

Messfehler auftreten. Außerdem ist die Kenntnis der Abhängigkeit zwischen gemessener Größe und externem Kriterium erforderlich.

Im einfachsten Fall messen externes Kriterium und das auf seine Validität zu überprüfende Messinstrument das Gleiche. In diesem Fall spricht man von der *Übereinstimmungsvalidität* (vgl. SCHNELL et al., 2005, S. 155). Die Korrelation zwischen den Messergebnissen des interessierenden Messinstruments und des externen Kriteriums ist dann ein Maß für die Validität der Messung:

$$V_X = r_{XY} \tag{3.40}$$

$V_X$: *Validität des Messinstruments X*

$r_{xy}$: *Korrelation zwischen Messinstrument X und externem Kriterium Y*

In Abbildung 3.27 soll diese Vorgehensweise bei der Beurteilung eines einzelnen Items exemplarisch dargestellt werden:

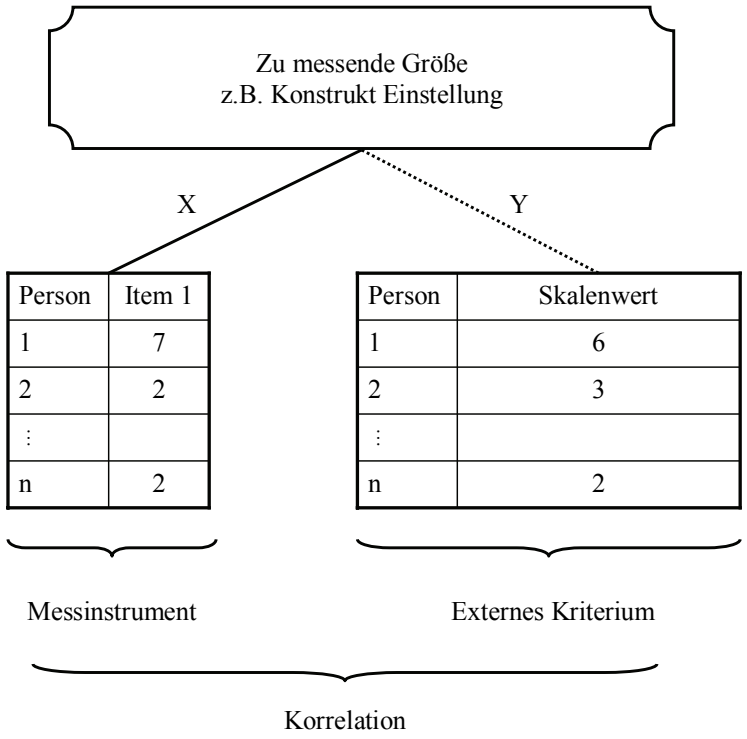


**Abbildung 3.27: Vorgehensweise bei der Bestimmung der Übereinstimmungsvalidität**

Die *Vorhersagevalidität* ist ein weiterer Sonderfall der Kriteriumsvalidität. Das externe Kriterium wird hier erst zu einem späteren Zeitpunkt erhoben. Die

Vorhersagevalidität misst demnach, ob durch die Messung eines Konstruktes in der Gegenwart, ein durch andere Messinstrumente in der Zukunft gemessenes externes Kriterium, prognostiziert werden kann (vgl. HAMMANN, ERICHSON, 2000, S. 77).

$$V_X = r_{X_{t_1} Y_{t_2}} \tag{3.41}$$

Ein Vergleich der Vorgehensweise bei der Vorhersagevalidität (siehe Abbildung 3.28) mit der Bestimmung der Reliabilität mit Hilfe der Test-Retest-Methode (Seite 82f.) zeigt, dass diese beiden Verfahren weitestgehend identisch sind. Der zentrale Unterschied der beiden Methoden besteht darin, dass bei der Test-Retest-Methode zur Ermittlung der Reliabilität der Messung im Zeitpunkt $t_2$ die gleichen Messinstrumente verwendet werden, während bei der Vorhersagevalidität im Zeitpunkt $t_2$ ein anderes Messinstrument zum Einsatz kommt (vgl. SCHNELL et al., 2005, S. 155).

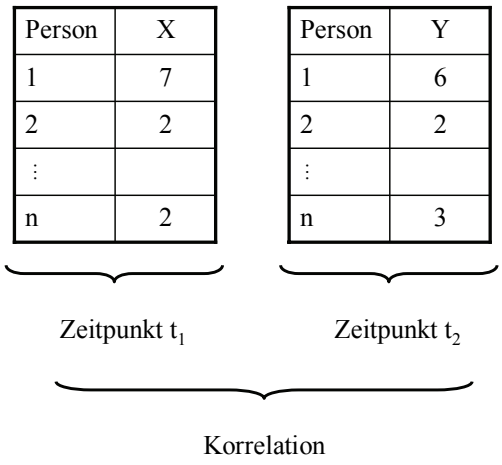


**Abbildung 3.28: Vorgehensweise bei der Messung der Vorhersagevalidität**

Die Kriteriumsvalidität ist ein geeignetes Maß zur Beurteilung der Qualität von Messergebnissen. Allerdings hängt die Anwendbarkeit sehr stark von der Existenz einer validen Messung des externen Kriteriums ab. Falls kein externes Kriterium zur Bewertung herangezogen werden kann, muss die Bewertung der Validität der Messergebnisse auf andere Weise geschehen.

Eine abgeschwächte Form der Übereinstimmungsvalidität prüft die Ergebnisse lediglich auf ihre Plausibilität. Diese Vorgehensweise erfordert kein exakt

gemessenes externes Kriterium. Anstelle dessen kann das externe Kriterium beispielsweise die gewählte Partei bei der letzten Bundestagswahl sein. Dieses sollte dann keinen Widerspruch zur Einstellung der Person zu Bundeswehr-einsätzen im Ausland (zu messende Größe) ergeben.

### 3.4.3.2 Konstruktvalidität

Bei der Konstruktvalidität müssen für ein Konstrukt Aussagen über Zusammenhänge zu anderen Konstrukten vorliegen. Wenn diese theoretisch abgeleiteten Beziehungen empirisch bestätigt werden, spricht man von der Konstruktvalidität (vgl. SCHNELL et al., 2005, S. 156). Dazu wird auf die in Kapitel 3.3.3.5 vorgestellten Strukturgleichungsmodelle zurückgegriffen.

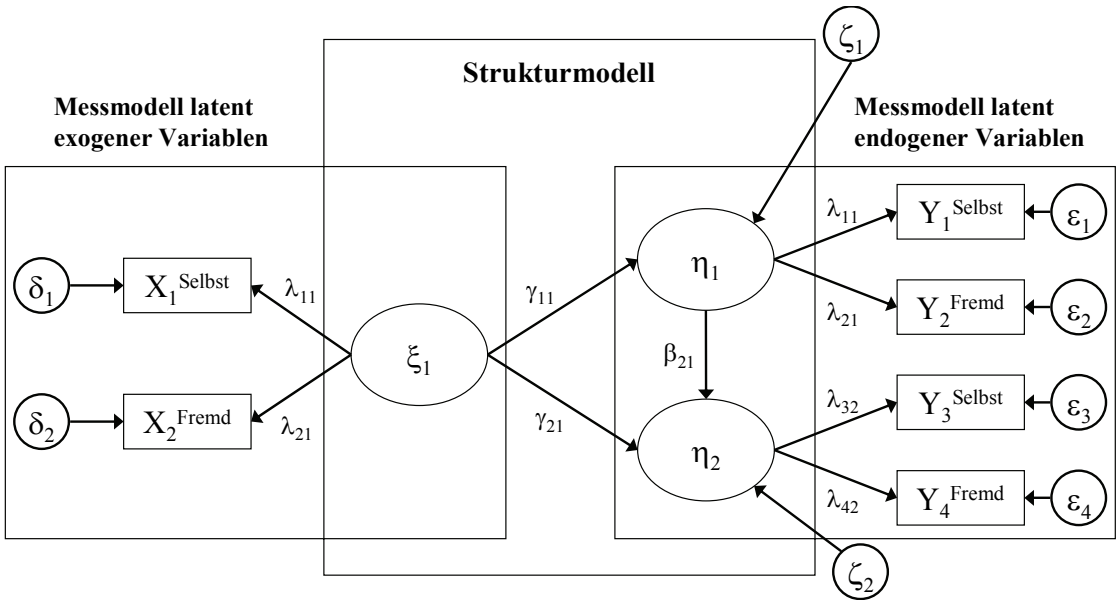


**Abbildung 3.29: Validitätsprüfung mit Hilfe der Multi-Trait-Multi-Method-Messung (vgl. EID et al., 2006, S. 285)**

Im ersten Schritt müssen für ein interessierendes Konstrukt ($\xi_1$ in Abbildung 3.29) und die zur Kontrolle verwendeten Konstrukte ($\eta_1$ und $\eta_2$ in Abbildung 3.29) die theoretischen Beziehungen festgestellt werden. Die Konstrukte werden mit Hilfe verschiedener Indikatoren gemessen (X, Y in Abbildung 3.29). Zum Schluss müssen die beobachteten Beziehungen zwischen den Konstrukten ($\gamma_{11}$, $\gamma_{21}$ in Abbildung 3.29) dahingehend überprüft werden, ob sie den theoretischen

Beziehungen entsprechen (vgl. SCHNELL et al., 2005, S. 156). Stimmen die empirisch beobachtbaren Beziehungen mit den theoretisch abgeleiteten Beziehungen nicht überein, kann dies drei Ursachen haben (vgl. SCHNELL et al., 2005, S. 157):

- Die Messung des interessierenden Konstruktes ($\xi_1$) ist nicht valide.
- Die aufgestellten Hypothesen über den Zusammenhang der Konstrukte ($\xi_1$, $\eta_1$, $\eta_2$) sind falsch.
- Das zu testende Konstrukt ist valide gemessen, aber die zur Beurteilung herangezogenen Konstrukte ($\eta_1$, $\eta_2$) sind ungenau gemessen.

Neben der Aussage über den Zusammenhang mit anderen Konstrukten müssen verschiedene Indikatoren desselben Konstrukts ($X_1$ und $X_2$) sehr stark korrelieren (vgl. HAMMANN, ERICHSON, 2000, S. 77). Darüber hinaus wird gefordert, dass diese Korrelation höher ist als die Korrelation zwischen Items unterschiedlicher Konstrukte (z.B. $X_1$ und $Y_1$). Schließlich sollen auch noch verschiedene Items desselben Konstrukts jeweils ähnliche Korrelationen zu Items anderer Konstrukte vorweisen (z.B. $r_{X_1Y_1} \approx r_{X_2Y_1}$) (vgl. SCHNELL et al., 2005, S. 151).

Damit eine Messung als valide angesehen wird, müssen alle vier Voraussetzungen erfüllt sein:

- $\gamma_{11}$, $\gamma_{21}$ entsprechen den theoretisch abgeleiteten Beziehungen
- $r_{X_1X_2}$ ist sehr hoch
- $r_{X_iX_j} > r_{X_iY_j}$ für alle i,j
- $r_{X_iY_j} \approx r_{X_iY_j}$ für alle i,j

Als besonders geeignet zur Überprüfung der Validität gelten Multi-Trait-Multi-Method-Modelle (vgl. SCHMITT, 2006, S. 17). Als Besonderheit weisen diese Messmodelle auf, dass bei der Bestimmung der latenten endogenen und exogenen Variablen (X,Y) unterschiedliche Messmethoden zum Einsatz kommen (vgl. EID et al., 2006, S. 284). Denkbar wären hier durch Fremdeinschätzung und Selbsteinschätzung erhobene Indikatoren wie sie in Abbildung 3.29 zusätzlich integriert sind (vgl. EID et al., 2006, S. 285).

## 3.4.4 Weitere Kriterien zur Beurteilung der Messqualität

Da insbesondere die Überprüfung der Validität in der Praxis große Probleme bereitet, haben sich weitere Verfahren zur Bestimmung der Messqualität herauskristallisiert. Diese Methoden erheben nicht den Anspruch, die Validität oder Reliabilität exakt zu bestimmen. Sie geben lediglich Anhaltspunkte für die Güte einer Messung. Es haben sich vor allem zwei Verfahren etabliert:

- Diskriminierungsfähigkeit der Indikatoren (Kapitel 3.4.4.1)
- Selbsteinschätzung durch die Probanden (Kapitel 3.4.4.2)

In der empirischen Studie in Kapitel 6 werden zusätzlich die bei einer Messung aufgetretenen Fehler mit Hilfe von Boxplots veranschaulicht. Daher soll diese Vorgehensweise hier kurz thematisiert werden (Kapitel 3.4.4.3), auch wenn in der entsprechenden Literatur nicht auf Boxplots zurückgegriffen wird.

### 3.4.4.1 Diskriminierungsfähigkeit der Indikatoren

Der Ausgangspunkt für die Ermittlung der Diskriminierungsfähigkeit ist die Messung eines bestimmten Konstrukts durch mehrere verschiedene Indikatoren. Mit Hilfe dieser Indikatoren lässt sich ein aggregierter Skalenwert für das Konstrukt berechnen (Gesamtwert), beispielsweise durch Mittelwertbildung. Auf Grundlage dieses Gesamtwertes werden zwei Personengruppen gebildet, eine Gruppe mit sehr hohen und eine Gruppe mit sehr niedrigen Werten bezogen auf das gesamte Konstrukt. Üblicherweise umfassen diese Gruppen je 25% der Versuchspersonen. In Abbildung 3.30 ist diese Gruppeneinteilung für die Messung eines Konstrukts mit drei Items vorgenommen:

| Person | Item 1 | Item 2 | Item 3 |
|---|---|---|---|
| 1 | 7 | 6 | 7 |
| 2 | 6 | 6 | 6 |
| 3 | 5 | 4 | 7 |
| 4 | 5 | 5 | 5 |
| 5 | 3 | 4 | 5 |
| 6 | 4 | 5 | 2 |
| 7 | 2 | 3 | 1 |
| 8 | 1 | 2 | 1 |

| Person | Gesamtwert | |
|---|---|---|
| 1 | 6,7 | Gruppe 1 |
| 2 | 6 | |
| 3 | 5,3 | |
| 4 | 5 | |
| 5 | 4 | |
| 6 | 3,7 | |
| 7 | 2 | Gruppe 2 |
| 8 | 1,3 | |

**Abbildung 3.30: Gruppeneinteilung zur Bestimmung der Diskriminierungsfähigkeit**

Für jedes Item kann anschließend der Mittelwert innerhalb der Personengruppen 1 und 2 bestimmt werden. Die Diskriminierungsfähigkeit eines Items entspricht dem Mittelwertunterschied zwischen den beiden Personengruppen. Bei einem sehr großen Mittelwertunterschied hat das betreffende Item eine sehr starke Diskriminierungsfähigkeit. In Abbildung 3.31 ist die Bestimmung der Diskriminierungsfähigkeit für Item 1 schematisch dargestellt:

| Person | Gesamtwert | Item 1 | |
|---|---|---|---|
| 1 | 6,7 | 7 | Gruppe 1 |
| 2 | 6 | 6 | |
| 3 | 5,3 | 5 | |
| 4 | 5 | 5 | |
| 5 | 4 | 3 | |
| 6 | 3,7 | 4 | |
| 7 | 2 | 2 | Gruppe 2 |
| 8 | 1,3 | 1 | |

| Gruppe | Mittelwert (Item 1) |
|---|---|
| 1 | 6,5 |
| 2 | 1,5 |

**Abbildung 3.31: Bestimmung der Diskriminierungsfähigkeit**

Nur wenn alle Items gemeinsam den betreffenden Sachverhalt geeignet abbilden, kann die Diskriminierungsfähigkeit einzelne Items zu identifizieren, die das entsprechende Konstrukt unzureichend messen. Die Diskriminierungsfähigkeit liefert jedoch keine Aussage darüber, ob alle Indikatoren gemeinsam überhaupt

den Sachverhalt geeignet messen. Daher kann die Diskriminierungsfähigkeit lediglich als Hilfsgröße bei der Beurteilung der Qualität der Messergebnisse betrachtet werden.

### 3.4.4.2 Selbsteinschätzung durch die Probanden

Die in dieser Arbeit bisher vorgestellten Ansätze versuchen alle ein objektives Maß zur Beurteilung der Qualität der Messergebnisse zu bestimmen. Diese Verfahren können jedoch nicht immer eingesetzt werden, da die erforderlichen Voraussetzungen (z.B. die valide Messung eines externen Kriteriums) zur Berechnung der Größen nicht zwangsläufig erfüllt sind. Deshalb dient neben diesen objektiven Methoden die subjektive Wahrnehmung der Versuchspersonen als zusätzlicher Bewertungsmaßstab. Die Probanden werden dazu befragt, ob ihrer Meinung nach die betreffenden Indikatoren zur Messung des Konstrukts geeignet sind. Bei der Beurteilung der Eignung eines Messinstruments durch die Probanden haben sich vor allem die folgenden Fragestellungen als zweckmäßig herauskristallisiert (vgl. PRESTON, COLMAN, 2000, S. 10):

- Wie einfach war für Sie der Umgang mit den Fragen?
- Wie schnell konnten Sie die Fragen beantworten?
- Haben Ihnen die Fragen die Möglichkeit gegeben, ihre Meinung exakt zu auszudrücken?

Ein für den Probanden einfach zu handhabendes Messinstrument ist vorteilhaft, da dadurch die Gefahr minimiert wird, dass Fehler auf Grund von Missverständnissen und Fehlinterpretationen entstehen. Die Schnelligkeit bei der Beantwortung ist ein positiver Aspekt, da die Aufmerksamkeit der Versuchspersonen bei einer Befragung im Zeitverlauf abnimmt (vgl. PRESTON, COLMAN, 2000, S. 10). Zur Schnelligkeit bei der Beantwortung steht die Möglichkeit, exakte Antworten zu formulieren, häufig im Widerspruch. Die Exaktheit der Antworten ist daher das dritte relevante Kriterium. Ein Beispiel dafür ist die Frage nach dem Einkommen einer Versuchsperson mit Hilfe von Einkommensintervallen. Verwendet man eine sehr grobe Einteilung in lediglich drei Gruppen (0-1.000, 1.000-2.000, über 2.000 Euro) ist die Frage für den Probanden deutlich schneller zu beantworten, als wenn insgesamt 30 Kategorien zur Verfügung

stehen. Dafür kann der Befragte bei einer Einteilung in 30 Gruppen eine genauere Antwort geben.

Die Schwachstellen der Selbsteinschätzung der Probanden sind offensichtlich. Bei der Messung eines Konstrukts sind Probanden kaum in der Lage, die Eignung der verschiedenen Indikatoren zu bewerten. Daher kommt die Selbsteinschätzung vor allem dann zum Einsatz, wenn verschiedene Indikatoren miteinander verglichen werden, beispielsweise bei der Frage, wie viele Antwortkategorien eines Items am besten sind.

Allerdings hat die Selbsteinschätzung auch beim Vergleich von Indikatoren Nachteile. So fällt es den Versuchspersonen wahrscheinlich schwer, zuzugeben, dass sie mit den gestellten Fragen überfordert waren. Des Weiteren ist zu bezweifeln, dass Versuchspersonen stets beurteilen können, ob ein Indikator ihnen die Möglichkeit zur exakten Antwort eingeräumt hat. Insgesamt sollte die Selbsteinschätzung daher eher zur zusätzlichen Absicherung der Ergebnisse und nicht als selbständiges Maß bei der Beurteilung von Messergebnissen herangezogen werden.

#### 3.4.4.3 Boxplots zur graphischen Veranschaulichung

Für die graphische Veranschaulichung der Eignung eines Indikators zur Messung eines Sachverhalts ist die Existenz eines – auf mindestens ordinalem Skalenniveau gemessenen – externen Kriteriums zwingend erforderlich. Bei den folgenden Ausführungen soll daher davon ausgegangen werden, dass die Einstellung einer Person zur Kirche bekannt ist (mit einem Wertebereich von 1-100). Diese dient als externes Kriterium ($y$). Als einzelner zu überprüfender Indikator ($x$) wird die Zustimmung zu folgender Aussage herangezogen: „Die Erhebung der Kirchensteuer in Deutschland ist sinnvoll.“ Folgende Antwortkategorien stehen den Probanden zur Verfügung:

- Stimme nicht zu
- Neutral
- Stimme zu

Mit Hilfe dieser Antwortkategorien können die Probanden in drei verschiedene Gruppen eingeteilt werden. Eine Gruppe von Personen, welche die Aussage ablehnen (Gruppe 1). Eine Personengruppe, die der Aussage neutral gegenübersteht (Gruppe 2) und jene Probanden, die der Aussage zustimmen (Gruppe 3). Abbildung 3.32 verdeutlicht diese Einteilung:

| Person | Einstellung zur Kirche (y) | x | |
|---|---|---|---|
| 1 | 7 | Stimme nicht zu | Gruppe 1 |
| 2 | 21 | Stimme nicht zu | |
| 3 | 71 | Stimme nicht zu | |
| ⋮ | ⋮ | ⋮ | |
| i | 29 | Neutral | Gruppe 2 |
| i+1 | 62 | Neutral | |
| ⋮ | ⋮ | ⋮ | |
| j | 34 | Stimme zu | Gruppe 3 |
| j+1 | 82 | Stimme zu | |
| ⋮ | ⋮ | ⋮ | |
| n | 96 | Stimme zu | |

**Abbildung 3.32: Gruppeneinteilung auf Basis des einzelnen Indikators**

Wenn der Indikator (Kirchensteuer ist sinnvoll) die Einstellung zur Kirche fehlerfrei misst, muss folgendes eintreten:

- Die Personen in Gruppe 1 (Stimme nicht zu) sollten die niedrigsten Werte beim externen Kriterium (Einstellung zur Kirche) aufweisen.
- In Personengruppe 2 (Neutral) sollten sich jene Personen befinden, die eine neutrale Einstellung zur Kirche haben (und daher einen mittleren Wert).
- Die Personengruppe 3 (Stimme zu) sollte schließlich die höchsten Werte bei der Einstellung zur Kirche vorweisen.

Es ist ein Indiz für eine geringe Qualität des Indikators, wenn viele Personen aus Personengruppe 1 einen höheren Einstellungswert haben als die Personen aus den Gruppen 2 und 3. Dies gilt ebenso für Personen aus Gruppe 2, die einen höheren Einstellungswert als Personen der Gruppe 3 haben. Diese Überschneidungen im Einstellungswert zwischen den unterschiedlichen

Personengruppen lassen sich mit Hilfe eines Boxplots graphisch darstellen (siehe Abbildung 3.33):

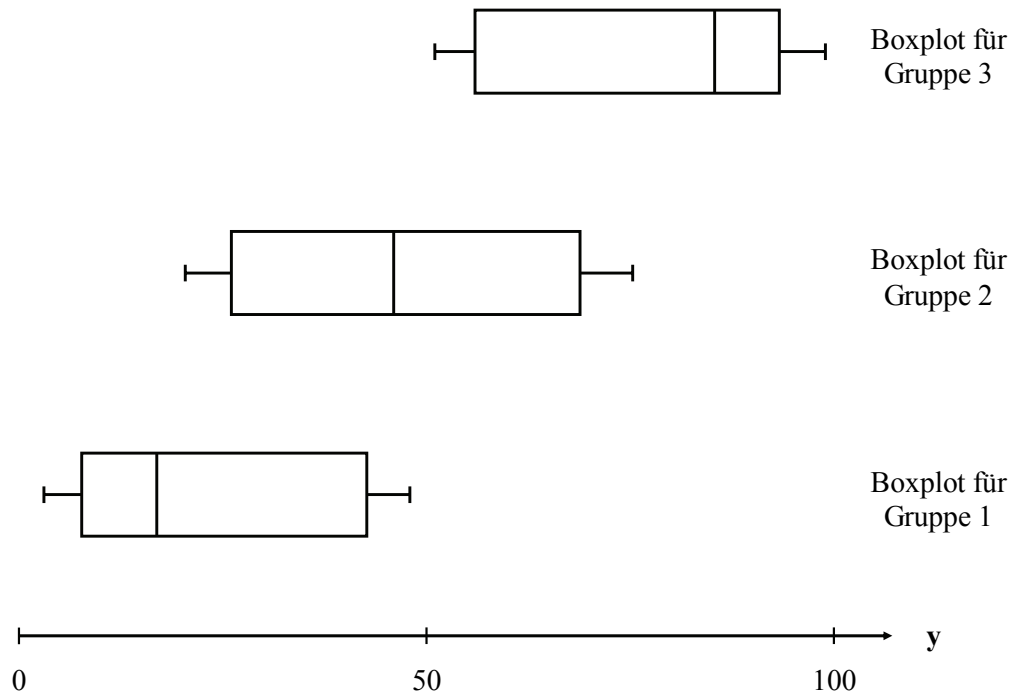


**Abbildung 3.33: Beurteilung der Qualität eines Indikators mit Hilfe eines Boxplots**

Die Boxplots verdeutlichen auf sehr einfache Weise, wie stark die Überschneidungen zwischen den Personengruppen sind. Allerdings liefern sie kein konkretes Maß für die Qualität einer Messung. Kritisch ist außerdem, dass Messwerte für das externe Kriterium *y* zwingend erforderlich sind. Falls ein solches externes Kriterium jedoch vorliegt, können die Boxplots als zusätzliches Instrument zur Visualisierung der Qualität einer Messung genutzt werden.

## 3.5 Zusammenfassung und Einordnung

In diesem Kapitel wurde ausführlich diskutiert, wie eine Messung im wirtschafts- und sozialwissenschaftlichen Bereich durchgeführt werden kann. Darüber hinaus wurde verdeutlicht, welche Schwierigkeiten bei Messungen in diesem Themenbereich auftreten und wie die Qualität einer Messung beurteilt werden kann.

Als Ergebnis dieses Kapitels bleibt festzuhalten, dass häufig auf das Instrument der Befragung zur Datenerhebung zurückgegriffen wird. Ein zentraler Bestandteil vieler Befragungen sind Ratingskalen. Diese lösen zwar nicht das zentrale Messproblem im wirtschafts- und sozialwissenschaftlichen Bereich, dass die interessierende Größe häufig nicht direkt beobachtbar ist, sondern nur indirekt mittels Indikatoren bestimmt werden kann. Aber sie bilden die Grundlage für viele ein- und mehrdimensionale Skalierungsverfahren, die dieses Phänomen berücksichtigen.

Bei der Bewertung der Qualität einer Messung muss die Frage beantwortet werden, wie gut gemessen wird, was gemessen werden soll. Kennzahlen, welche die Qualität einer Messung bestimmen, sind die Reliabilität und Validität. Die Reliabilität dient dabei der Beurteilung, inwieweit die Messung frei von zufälligen Fehlern ist. Mit Hilfe der Validität wird überprüft, ob bei der Messung der Sachverhalt gemessen wird, der gemessen werden soll. Daneben existieren noch weitere Kriterien, falls die Reliabilität und Validität nicht bestimmbar sind.

Wie gut ein bestimmter Sachverhalt durch ein Skalierungsverfahren gemessen wird, muss im Einzelfall entschieden werden. Wie bereits erwähnt, beruhen jedoch nahezu alle Skalierungsverfahren auf der Messung von Indikatoren mittels Ratingskalen oder Paarvergleichen. Daher können Skalierungsverfahren nur dann brauchbare Ergebnisse liefern, wenn Ratingskalen oder Paarvergleiche zur Messung der Indikatoren geeignet sind. Da in dieser Arbeit der Fokus auf der Beurteilung der Ratingskala liegt, soll nochmals festgehalten werden, welche Skalierungsverfahren auf diese zurückgreifen:

- Likert-Skala
- Guttman-Skala
- Semantisches Differenzial
- Modell von FISHBEIN
- Modell von TROMMSDORFF

Bei der multidimensionalen Skalierung und Strukturgleichungsmodellen sind Ratingskalen nicht zwingend erforderlich. Dennoch liegen auch diesen Verfahren häufig Ratingskalen zu Grunde.

Dieses Kapitel hat gezeigt, wie die Messung selbst und die Beurteilung ihrer Qualität vorgenommen werden kann. Sie liefert allerdings keine Ansatzpunkte, warum bei einer Messung auch unerwünschte Effekte auftreten können, welche die Qualität der Messung negativ beeinflussen.

Im nun folgenden vierten Kapitel dieser Arbeit steht daher die Frage im Mittelpunkt, warum die Messergebnisse im wirtschafts- und sozialwissenschaftlichen Bereich häufig unbrauchbar sind. Diese Erkenntnisse ermöglichen es dann, die Qualität von Messergebnissen positiv zu beeinflussen.

# Kapitel 4

# Beantwortungseffekte und ihre Ursachen

Fehlerhafte oder nicht-exakte Antworten stellen eines der grundlegenden Probleme der Messung im wirtschaftswissenschaftlichen Bereich dar. Fehler treten nicht nur bei der Messung mittels einer Ratingskala, sondern allgemein bei jeder Befragung unabhängig von der gewählten Methode auf. Um nicht den Eindruck zu erwecken, die Probleme seien auf Ratingskalen beschränkt, sollen die Effekte hier allgemein für viele verschiedene Befragungsarten beschrieben werden. Die Gründe für fehlerhafte Antworten können dabei sehr vielfältig sein.

**Definition 4.1**

*Der Begriff* ***Response Effects (Beantwortungseffekte)*** *fasst alle Phänomene zusammen, die zu einer Antwortverzerrung oder zur Antwortverweigerung des Probanden aufgrund des kognitiven Beantwortungsprozesses bei der Befragung führen (vgl.* TOURANGEAU *et al., 2005, S. 2).*

Zur Herausarbeitung der genauen Ursachen dieser Beantwortungseffekte muss geklärt werden, welche Prozesse der Befragte bei einer Befragung durchläuft. In diesem Kapitel werden daher verschiedene psychologische Modelle vorgestellt, die erklären sollen, welche Prozesse dies im Einzelnen sind.

Ein Überblick über die Literatur zeigt, dass eine Vielzahl verschiedener Modelle existiert, welche jedoch teilweise im Widerspruch zueinander stehen. Daher wird in Kapitel 4.1 zunächst ein grundlegender Ablauf der psychologischen Prozesse vorgestellt, der im Einklang mit der Mehrzahl der existierenden Modelle steht. Dieses Basismodell soll als Grundlage für die weitere Arbeit dienen. Anschließend werden folgende konkrete Modelle vorgestellt:

- Das Modell von CANNELL, MILLER und OKSENBERG (1981) (Kapitel 4.2). Die Autoren befassen sich dabei insbesondere mit solchen Prozessen, die zu Beantwortungseffekten führen.

- Das Satisficing-Modell von KROSNICK und ALWIN (Kapitel 4.3) liefert eine Erklärung, weshalb je nach Motivation und intellektueller Fähigkeit einer Versuchsperson unterschiedliche Beantwortungseffekte entstehen.

- Das Two-Track-Modell von STRACK und MARTIN (Kapitel 4.4) verdeutlicht, wie Beantwortungseffekte durch den Fragebogenkontext entstehen können.

- Die bis dato ausführlichste Darstellung der kognitiven Prozesse einer Befragung liefern 2005 TOURANGEAU, RIPS und RASINSKI (siehe Kapitel 4.5).

## 4.1 Psychologische Prozesse bei der Befragung

In diesem Kapitel sollen jene psychologischen Prozesse vorgestellt werden, die in den Modellen nahezu aller Autoren vorkommen. Eine Übersicht über die bestehenden Modelle liefern TOURANGEAU, RIPS und RASINSKI (2005). Auf Grund der Heterogenität der Modelle kann lediglich eine grobe Einordnung der psychologischen Prozesse vorgenommen werden, welche als Grundlage für die weitere Arbeit jedoch sehr bedeutsam ist.

Die meisten Autoren gehen von vier Hauptkomponenten beim Beantwortungsprozess aus, die alle für Beantwortungseffekte verantwortlich sein können (vgl. TOURANGEAU et al., 2005, S. 8). Jede dieser vier Hauptkomponenten besteht wiederum aus mehreren kognitiven Prozessen, die nicht zwangsläufig alle zu durchlaufen sind (vgl. TOURANGEAU et al., 2005, S. 8). Abbildung 4.1 veranschaulicht diese kurz:

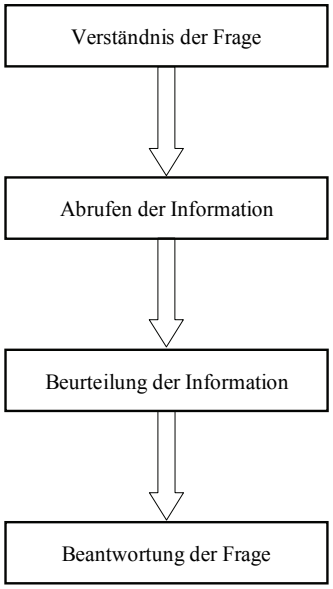


**Abbildung 4.1: Der Antwortprozess**

Am homogensten ist die Auffassung der Autoren bezüglich des Verständnisses der Frage. Deshalb sollen die Prozesse zum Verständnis der Frage im Rahmen von Kapitel 4.1.1 ausführlich erläutert werden.

Bei den nächsten beiden Schritten im Beantwortungsprozess – Abrufen und Beurteilung der Information – unterscheiden sich die verschiedenen Modelle grundlegend. Daher soll in Kapitel 4.1.2 lediglich kurz auf das Ergebnis dieser beiden Schritte eingegangen werden.

Kapitel 4.1.3 hat schließlich die Beantwortung der Frage und die dabei ablaufenden Prozesse zum Thema.

## 4.1.1 Verständnis der Frage

Das Verständnis der Frage stellt die Grundvoraussetzung für eine adäquate Beantwortung durch den Probanden dar. Verständnis kann auf zwei Wegen erreicht werden, die beide auf der gedanklichen Erfassung des gerade gehörten oder gelesenen Satzes basieren:

- Unmittelbares Verständnis (Intuition)
- Mittelbares Verständnis (Interpretation)

Es müssen dabei jeweils mehrere Prozesse durchlaufen werden, um zu unmittelbarem oder mittelbarem Verständnis zu gelangen (vgl. Abbildung 4.2):

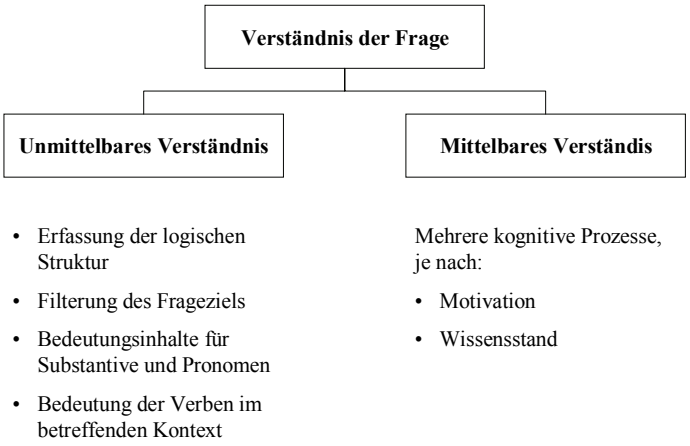


**Abbildung 4.2: Die kognitiven Prozesse beim Verständnis der Frage**

Zunächst müssen alle Personen die grammatikalische und logische Struktur eines Satzes erfassen (vgl. TOURANGEAU et al., 2005, S. 31). Dies entspricht dem *unmittelbaren Verständnis* der Frage. Dazu zählen folgende Prozesse (vgl. GRAESSER et al., 1996, S. 152):

- Erfassung der logischen Satzstruktur
- Filterung des Frageziels
- Verbindung der Substantive und Pronomen mit den im Gehirn gespeicherten Bedeutungsinhalten
- Zuordnung einer Bedeutung für alle Verben im betreffenden Kontext

Abgesehen von sprachlichen Problemen sollte dieser Teil bei allen Probanden zum gleichen Ergebnis führen. Der zweite (optionale) Teil wird durch die Verknüpfung der Frage mit bereits bestehendem Wissen der Auskunftsperson charakterisiert. Dieses *mittelbare Verständnis* (Frageinterpretation) variiert in Abhängigkeit

- des Wissensstandes und der
- Motivation der Probanden (vgl. TOURANGEAU et al., 2005, S. 31).

Inwiefern Motivation und Wissensstand einer Person die Beantwortung beeinflussen, wird beim Satisficing-Modell von KROSNICK und ALWIN ausführlich diskutiert. Daher sollen hier lediglich das unmittelbare Verständnis betreffende Probleme ausgeführt werden. Grundsätzlich können zwei Bereiche unterschieden werden:

- Unbestimmtheit
- Voraussetzung

Ein Problem entsteht, wenn bestimmte Wörter nicht von allen Personen einheitlich interpretiert werden, wie eine Studie von BELSON (1981) belegt. In der Studie wird gefragt, ob *Kinder* durch Gewalt im Fernsehen Schaden nehmen. Eine zusätzliche Befragung der Probanden zeigt, dass die Vorstellung über das Alter eines *Kindes* vom Kleinkind bis zum Teenager reicht (vgl. BELSON, 1981, S. 23). Diese Mehrdeutigkeit einzelner Wörter oder ganzer Sätze ist ein erstes zentrales Problem beim unmittelbaren Verständnis, das als *Unbestimmtheit* der Fragestellung bezeichnet wird. Das Phänomen der Unbestimmtheit von Ausdrücken ist selbstverständlich nicht auf die Fragestellung beschränkt. Bei geschlossenen Fragestellungen können diese Probleme auch bei den Antwortkategorien auftreten. In den meisten Fällen sind die Antwortalternativen verbal beschrieben, beispielsweise mit ´nie´, ´selten´, ´ab und zu´, ´häufig´ und ´immer´. Diese sprachlichen Formulierungen sind jedoch nicht exakt auf numerische Ausprägungen zurückzuführen, wie BRADBURN und SUDMAN bereits 1979 nachweisen (vgl. BRADBURN, SUDMAN, 1979, Kapitel 10). So folgt aus den Aussagen ´In Kalifornien kommt es *oft* zu Erdbeben´ und ´Die Person muss *oft* niesen´ nicht, dass die beiden Ereignisse gleich häufig eintreten (vgl. TOURANGEAU et al., 2005, S. 47). Bei diesem Beispiel ist die unterschiedliche Zuordnung der Zahlenwerte auf den jeweiligen Kontext zurückzuführen. Aber selbst bei identischer Fragestellung variiert die Zuordnung der tatsächlichen numerischen Werte zu den verbal formulierten Antwortalternativen zwischen den Probanden (vgl. MOXEY, SANFORD, 1993, S. 96). In einigen Fällen sind die Versuchspersonen gar nicht in der Lage, die exakte numerische Ausprägung zu bestimmen. Dann dient die sprachliche Antwortkategorie den Probanden lediglich als Möglichkeit, um ihre relative Position im Vergleich zu anderen zu dokumentieren (vgl. TOURANGEAU et al., 2005, S. 49). Die Aussage ´In Kalifornien kommt es *oft* zu Erdbeben´ entspricht damit nicht einer konkreten

Zahl von Erdbeben, sondern lediglich der Überzeugung, dass es in Kalifornien häufiger zu Erdbeben kommt als in anderen Gegenden.

Wenn durch die Fragestellung bestimmte Verhaltensmuster bzw. Einstellungen des Probanden unterstellt werden, spricht man vom Problem der *Voraussetzung*. So haben Männer bei der Frage, ob sie sich ´oft´, ´manchmal´, ´selten´ oder ´nie´ schlecht gefühlt haben, weil sie ihrer Frau untreu waren (vgl. AAKER, DAY, 1990, S. 187), Schwierigkeiten Stellung zu beziehen, wenn sie ihrer Frau immer treu waren. Die Fragestellung setzt voraus, dass die befragte Person tatsächlich untreu war. In der Studie gibt ein sehr hoher Anteil der Männer (85%) an, sich nie schlecht gefühlt zu haben (vgl. AAKER, DAY, 1990, S. 187). Dies liegt aber vielleicht daran, dass sie bisher immer treu waren.

### 4.1.2 Abrufen und Beurteilung relevanter Information

Wie bereits erwähnt existieren mehrere unterschiedliche Modelle zu den psychologischen Prozessen, die für Abrufen und Beurteilung relevanter Information erforderlich sind. Dabei beleuchten die verschiedenen Autoren großteils unterschiedliche Aspekte dieses Themas.

Weitestgehend einheitlich werden lediglich die psychologischen Prozesse bei der Zustimmung oder Ablehnung einer Aussage dargestellt. Deshalb soll dieser Teilbereich hier ausgeführt werden. Die Autoren gehen davon aus, dass Versuchspersonen wie bereits von THURSTONE (1927) postuliert, jede Frage mit Hilfe eines persönlichen *psychologischen Kontinuums* beantworten (siehe Kapitel 3.3.2.3).

Das Ergebnis des Abrufens und Beurteilens relevanter Information ist ein konkreter Zahlenwert, beispielsweise der Grad der Zustimmung zu einem Statement, auf dem psychologischen Kontinuum. Abbildung 4.3 veranschaulicht dies:

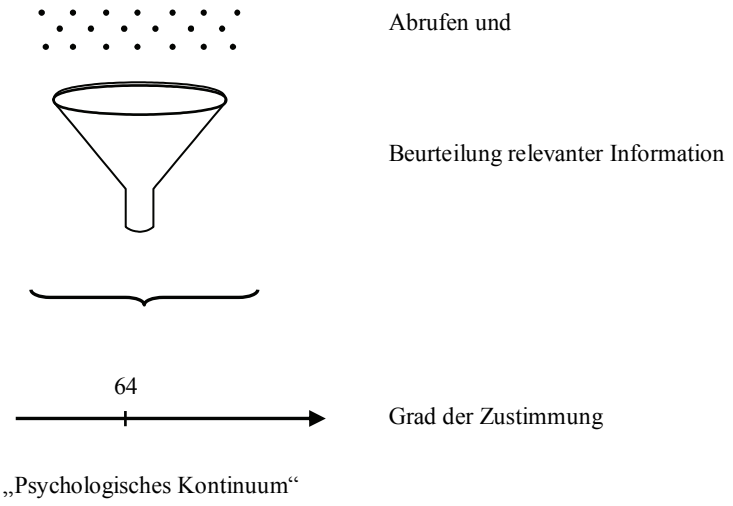


**Abbildung 4.3: Ergebnis des Abrufens und Beurteilens von Information**

Nachdem die Probanden die Fragestellung verstanden haben, müssen sie die zur Beantwortung erforderliche Information abrufen und beurteilen. Dieser Teil des Beantwortungsprozesses ist in Abbildung 4.3 durch den Trichter dargestellt. Am Ende dieses Prozesses bleibt als Ergebnis ein Grad der Zustimmung bzw. Ablehnung zur betreffenden Aussage.

Allerdings können die Probanden bei den meisten Fragestellungen diesen konkreten Zahlenwert nicht als Antwort angeben. In Abbildung 4.3 ist beispielsweise 64 das Ergebnis der Informationsbeurteilung. Dies wäre keine adäquate Antwort auf die Frage, ob die Person der Aussage zustimmt oder nicht. Vielmehr muss der Proband dieses Ergebnis in eine zur Frageformulierung und den vorhandenen Antwortmöglichkeiten passende Form transformieren.

## 4.1.3 Beantwortung der Frage

Neben der Transformation des Ergebnisses der Informationsbeurteilung in eine geeignete Form, stehen Probanden noch vor vielen weiteren Problemen bei der Beantwortung, etwa, weil das Ergebnis der Informationsbeurteilung nicht der Antwort entspricht, von der die Versuchsperson denkt, dass sie von ihr erwartet wird. Da die Hintergründe für eine Abänderung der Antwort in der Literatur

unterschiedlich erklärt sind, soll an dieser Stelle nur das Problem der Transformation der Werte thematisiert werden.

Zur Erklärung des Mappingproblems soll an den Beantwortungsprozess in Abbildung 4.3 angeknüpft werden. Das Ergebnis dieses Prozesses war ein konkreter Grad der Zustimmung. Angenommen, der Person steht zur Beantwortung eine Ratingskala mit 5 Antwortkategorien, von ´stimme zu´ bis ´stimme nicht zu´, zur Verfügung. Dann steht sie vor dem Problem, das Ergebnis des Informationsprozesses in eine der vorgegebenen Antwortkategorien abzubilden. Abbildung 4.4 zeigt, dass zwei Personen durch abweichende Mappingstrategien zu unterschiedlichen Ergebnissen kommen, obwohl die Informationsbewertung der beiden Personen übereingestimmt hat:

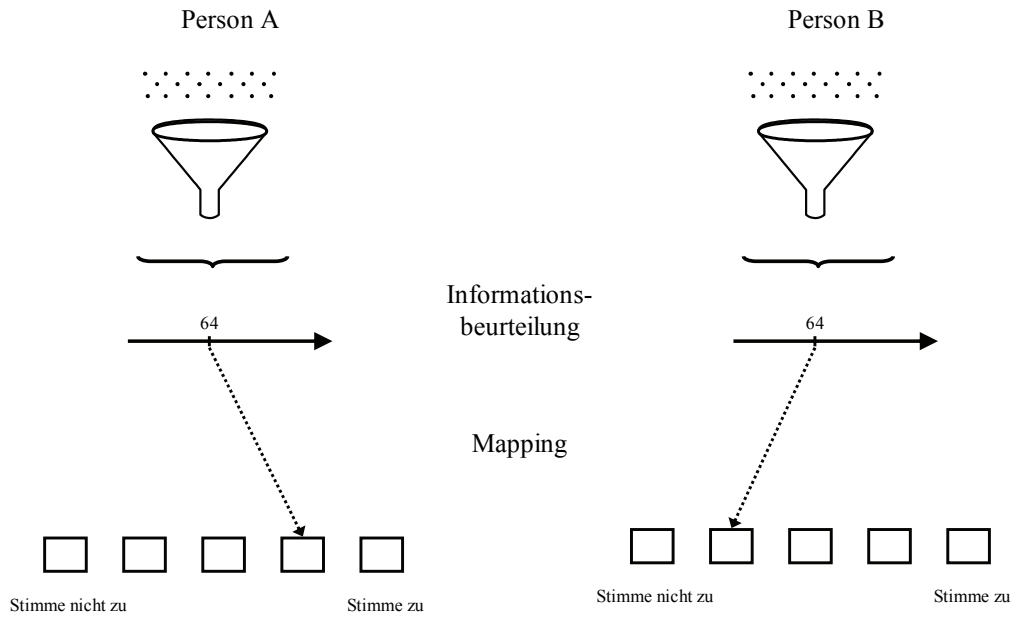


**Abbildung 4.4: Auswirkungen des Mappingprozesses**

Wie dieses Beispiel verdeutlicht, können Beantwortungseffekte auch bei gewissenhafter Beantwortung durch den Probanden entstehen. Im Allgemeinen kann nämlich nicht davon ausgegangen werden, dass alle Probanden identische Mappingstrategien verwenden.

Nach diesen einleitenden und eher allgemeinen Ausführungen sollen im Folgenden konkrete Modelle des Beantwortungsprozesses und insbesondere ihre Erklärung für Beantwortungseffekte dargestellt werden. Der in Abbildung 4.1

vorgestellte grundsätzliche Ablauf des Beantwortungsprozesses ist dabei bei allen Ansätzen erkennbar.

## 4.2 Das Modell von CANNELL et al.

In ihrem Modell befassen sich CANNELL, MILLER und OKSENBERG mit den Schritten, die ein Proband durchlaufen muss, um eine Frage richtig zu beantworten. Im Fokus der Autoren liegen jedoch insbesondere die Prozesse, die zur Antwortverweigerung oder Antwortverzerrung führen. In Abbildung 4.5 sind sowohl die Abläufe dargestellt, die zu einer richtigen Beantwortung führen, als auch die Gründe für Beantwortungseffekte (vgl. hierzu und zu den folgenden Ausführungen CANNELL et al., 1981, S. 389-437).

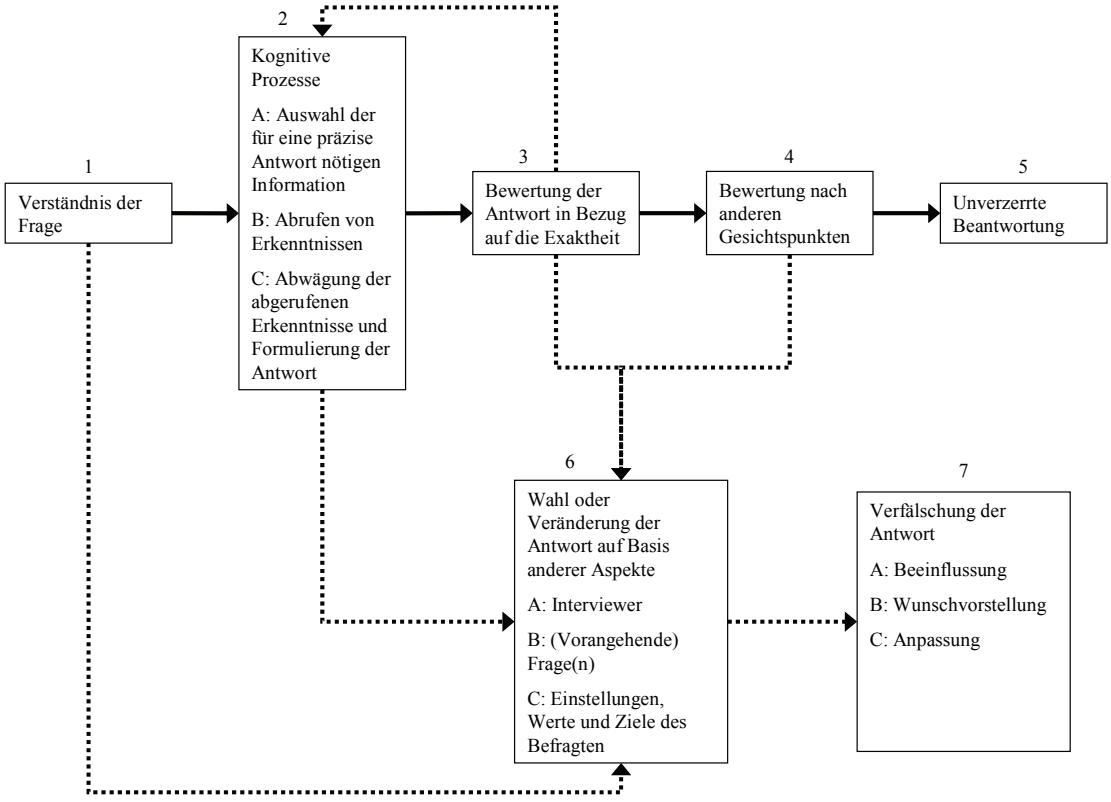


**Abbildung 4.5: Der Beantwortungsprozess nach CANNELL et al. (in Anlehnung an CANNELL et al., 1981, S. 393)**

Für eine korrekte Antwort muss die Versuchsperson folgende Prozesse durchlaufen:

1. Zunächst muss der Proband die Frage verstehen. Dazu zählt die Kenntnis des Vokabulars sowie die Klarheit der Begriffe und der gesamten Satzstruktur. Hat die Person die Frage verstanden und ist motiviert eine exakte Antwort zu geben, so folgt im zweiten Schritt der Informationsprozess.

2. Der Informationsprozess ist eine kognitive Abfolge, bestehend aus der Auswahl der Information sowie dem Abrufen und Abwägen von Erkenntnissen. Die Formulierung der Antwort erfolgt bereits in diesem frühen Stadium des Beantwortungsprozesses. Bevor der Befragte seine Antwort jedoch äußert, wird er sie zunächst für sich in den beiden nächsten Schritten bewerten.

3. In diesem Schritt wägt der Proband ab, ob seine Antwort aus Schritt 2 der Fragestellung gerecht wird. Falls die Antwort vom Befragten als nicht adäquat eingestuft wird, wiederholt er die Prozesse in Schritt 2.

4. In diesem Abschnitt bewertet die Auskunftsperson die Antwort erneut, allerdings vor dem Hintergrund, ob die Antwort mit der eigenen Meinung bzw. den eigenen Zielen übereinstimmt.

5. Falls zwischen der Antwort und den persönlichen Werten und Zielen des Befragten kein Widerspruch besteht, führt dies zur unverzerrten Beantwortung.

Neben diesem idealtypischen Ablauf zur korrekten Beantwortung der Frage existieren die in der unteren Hälfte abgebildeten Einflüsse, die eine ungenaue Antwort zur Folge haben (Schritte 6,7).

6. Die in diesem Schritt dargestellten Störgrößen beziehen sich auf den Interviewer, die vorangehenden Fragen, aber auch die Einstellungen, Werte und Ziele der Auskunftsperson.

Falls eine Person die Frage nicht verstanden hat (Schritt 1), oder falls ihr die Schritte 2-4 als zu aufwendig erscheinen, geht sie zu Punkt 6 über und lässt die für eine akkurate Beantwortung nötigen Teilaspekte außer Acht. Die Gründe für den Übergang auf Schritt 6 sind vielschichtig. In Schritt 2 entstehen

Beantwortungseffekte durch fehlerhafte kognitive Prozesse, z.B. weil vorangehende Fragen die Erinnerung der Versuchsperson beeinflussen. Es handelt sich also um von den Probanden unabsichtlich herbeigeführte Beantwortungseffekte. Bei Schritt 3 und 4 erfolgt die Bewertung der Antwort. Falls diese als nicht geeignet betrachtet wird, kann mangels Motivation auf das erneute Durchlaufen von Schritt 2 verzichtet werden und direkt auf Schritt 6 übergegangen werden.

7. Wenn eine Versuchperson zu Schritt 6 übergegangen ist, hat dies eine verfälschte Antwort (Schritt 7) zur Folge. Eine verfälschte Antwort entsteht durch Beeinflussung, der Wunschvorstellung des Probanden und der Anpassung der Antwort.

Eine *Beeinflussung* des Probanden entspringt der Fragesituation. Es ist häufig keine vom Probanden bewusst herbeigeführte Antwortverfälschung. Die Auskunftsperson wird durch den Interviewer und vorangehende Fragen unbewusst beeinflusst, was insbesondere in Schritt 2 zu Fehlern führt.

Die *Wunschvorstellung* beschreibt dagegen Veränderungen der Antwort, weil die Person ihrem individuellen Ideal entsprechen möchte. Die Bewertung der Antwort in Schritt 4 kann z.B. eine Diskrepanz zwischen aufrichtiger Antwort und persönlicher Idealvorstellung aufzeigen. Die Folge daraus kann eine unehrliche Antwort sein, um der eigenen Wunschvorstellung zu entsprechen.

Mit *Anpassung* ist gemeint, dass die Auskunftsperson versucht, ihre Antwort dem anzupassen, was sie als gesellschaftlich opportun erachtet. Liegt die Antwort außerhalb eines sozial erwünschten Bereichs, fürchtet die Person negative Sanktionen. Personen, die im Bereich der negativen Sanktionen liegen, tendieren dazu, eine unwahre oder keine Antwort auf die Frage zu geben. In Abbildung 4.6 sind dies die Personen A und D.

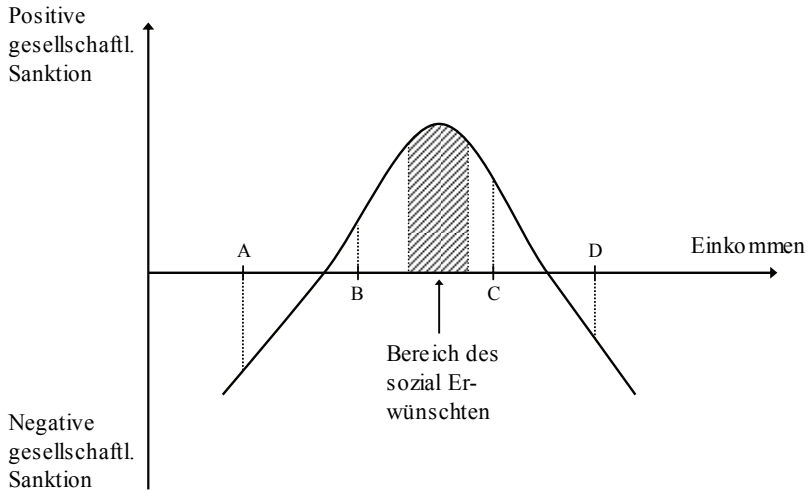


**Abbildung 4.6: Antwortverweigerung in Abhängigkeit der Merkmalsausprägung (in Anlehnung an HOLM, 1975, S. 83)**

## 4.3 Das Satisficing-Modell von KROSNICK und ALWIN

Das Grundprinzip des Ansatzes von CANNELL, MILLER und OKSENBERG findet sich auch beim Satisficing-Modell von KROSNICK und ALWIN wieder. In ihrer empirischen Studie verwenden die Autoren allerdings keine Ratingskalen, sondern Auswahlfragen. Da das Satisficing-Modell in der Psychologie jedoch sehr bedeutsam ist und die grundsätzlichen Ergebnisse unabhängig von der Fragetechnik sind, soll das Modell hier erwähnt werden.

KROSNICK und ALWIN unterstellen wie CANNELL, MILLER und OKSENBERG zwei unterschiedliche Wege, auf denen Probanden zur Beantwortung einer Frage gelangen können (vgl. hierzu und zum folgenden Abschnitt KROSNICK, ALWIN, 1987, S. 201-219). Allerdings unterscheiden KROSNICK und ALWIN dabei vor allem die Motivation und die intellektuellen Fähigkeiten der Befragten. Sie stellen die These auf, dass eine motivierte und intellektuell fähige Person eine Frage *wahrheitsgemäß* beantworten will. Unmotivierte oder intellektuell überforderte Personen versuchen dagegen nur, eine *akzeptable bzw. plausible* Antwort zu finden. Dies entspricht dem in der Psychologie weit verbreiteten Satisficing-Prinzip, wonach Personen generell nicht nach optimalen, sondern

lediglich nach akzeptablen Problemlösungen suchen (vgl. SIMON, 1957, S. 204f.).

Um ihre These empirisch zu unterstützen, teilen die Verfasser die Personen nach ihren intellektuellen Fähigkeiten in zwei Gruppen ein. Den Personen in beiden Gruppen wird eine Liste mit 13 positiven Eigenschaften eines Kindes vorgelegt, etwa gute Manieren, Ehrlichkeit oder Erfolg. Anschließend müssen die Personen angeben, welche 3 Eigenschaften in ihren Augen die Wichtigsten bei einem Kind seien. Dabei variiert die Reihenfolge der 13 positiven Eigenschaften. Personen, die bloß nach einer *plausiblen* Antwort suchen, werden vor allem die Antworten am Beginn der Liste auswählen, unabhängig welche Eigenschaft dies ist. Dieses Phänomen ist in der Fragebogenpsychologie als *Reihenfolgeeffekt* bekannt.

Falls kein Reihenfolgeeffekt zu beobachten ist, die Antworten der Personen also unabhängig von der Reihenfolge sind, ist dies ein Hinweis, dass die Probanden die Fragen *wahrheitsgemäß* beantworten. In ihrer empirischen Studie können KROSNICK und ALWIN belegen, dass in der Personengruppe mit größeren intellektuellen Fähigkeiten deutlich weniger Reihenfolgeeffekte auftreten, als in der Gruppe mit geringeren intellektuellen Fähigkeiten. Daraus folgern sie, dass Personen mit größeren intellektuellen Fähigkeiten nach optimalen Lösungen suchen, während Personen mit geringeren intellektuellen Fähigkeiten nur nach passablen Lösungen suchen.

Die Existenz von Reihenfolgeeffekten ist in vielen Studien nachgewiesen und daher allgemein anerkannt. Es erscheint ebenfalls sinnvoll, dies mit der Bereitschaft der Versuchspersonen zu begründen, *wahrheitsgemäße* Antworten und nicht lediglich *akzeptable* Antworten zu geben. Ob dafür allerdings die intellektuellen Fähigkeiten entscheidend sind, ist fragwürdig. Vielmehr kann die Motivation der Probanden eine entscheidende Rolle spielen. Motivierte Versuchspersonen antworten deutlich überlegter und geben dadurch eher wahre Antworten. Unmotivierte Probanden wollen dagegen nicht lange nachdenken. Sie sind deshalb zufrieden, wenn ihre Antworten plausibel scheinen.

## 4.4 Das Two-Track-Modell von STRACK und MARTIN

Bei STRACK und MARTIN steht die Einbettung einer Frage in den gesamten Kontext des Fragebogens im Mittelpunkt der Betrachtung (vgl. hierzu und zum folgenden Absatz STRACK, MARTIN, 1987, S. 123-148). Dazu stellen die Autoren jedoch auch ein Modell der kognitiven Prozesse bei der Fragebeantwortung auf. Wie der Name des Modells schon vermuten lässt, basiert auch dieser Ansatz auf dem Gedanken, dass Probanden auf zwei unterschiedlichen Wegen zur Beantwortung gelangen können, wie Abbildung 4.7 verdeutlicht. Im Gegensatz zu den bisher behandelten Ansätzen erklären STRACK und MARTIN dadurch jedoch nicht, weshalb unexakte Antworten entstehen. Vielmehr versuchen sie durch ihre Unterscheidung die Bedeutung des gesamten Fragebogenkontexts herauszuarbeiten. Sie prägen daher die Begriffe *Zielfrage* und *Kontextfrage*. Bei der *Zielfrage* handelt es sich um die interessierende Fragestellung. *Kontextfragen* sind die der Zielfrage vorausgehenden Fragen.

Prinzipiell unterteilen sie den Beantwortungsprozess in vier Abschnitte (siehe Abbildung 4.7):

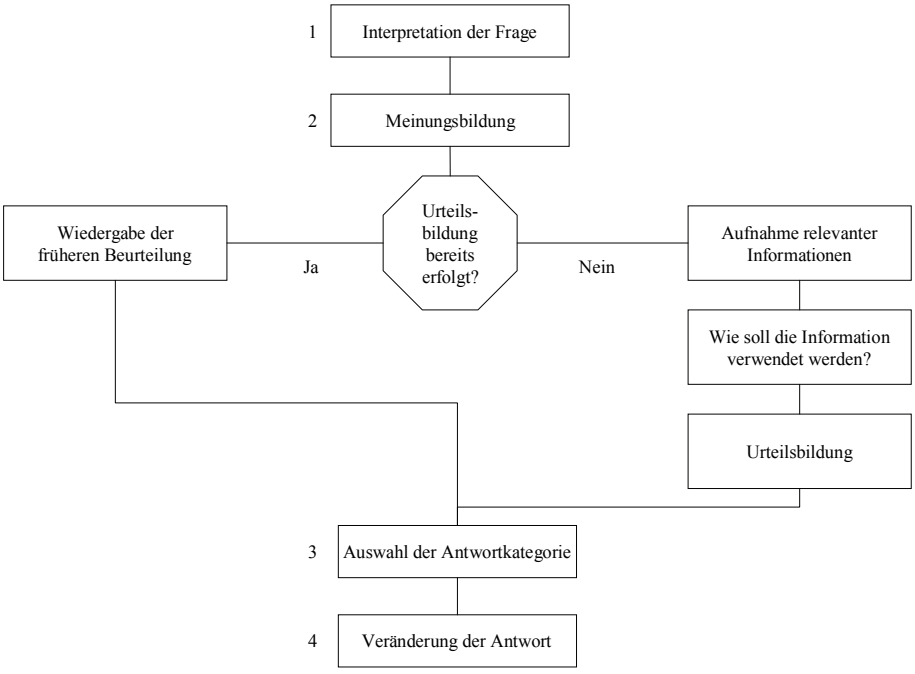


**Abbildung 4.7: Der Informationsprozess nach STRACK und MARTIN (in Anlehnung an STRACK, MARTIN, 1987, S. 125)**

1. Den ersten Abschnitt bezeichnen die Autoren als *Interpretation* der Frage. Dabei stehen den Probanden drei unterschiedliche Quellen zur Informationsgewinnung zur Verfügung:

   - Tatsächlicher Inhalt der Zielfrage
   - Antwortkategorien
   - Kontextfragen

   Neben der tatsächlichen Bedeutung der Frage, beurteilen die Versuchspersonen die Frage demnach zusätzlich nach den zur Verfügung gestellten Antwortkategorien und den Kontextfragen. Die Antwort-kategorien sind insbesondere dann von Bedeutung, wenn Begriffe in der Frage auf unterschiedliche Arten interpretiert werden können. In diesem Fall können auch Kontextfragen eine entscheidende Rolle spielen. Die Autoren geben als Beispiel die Frage nach den Leistungen von Ronald Reagan an. Je nach Kontext in den diese Frage eingebettet ist, zielt die Frage auf die Leistungen Ronald Reagans als Schauspieler oder als Politiker ab.

2. Im zweiten Abschnitt erfolgt die *Meinungsbildung* der Auskunftspersonen. Falls sie zu diesem Thema bereits feste Ansichten gebildet haben, so werden diese als Basis für die Beantwortung der Frage herangezogen. Dabei ist von sehr großer Bedeutung, dass die Personen nicht die ursprüngliche Information erneut bewerten, sondern sich lediglich an ihre Ansichten erinnern. Dieses Urteil wird nicht abgeändert, wenn es ursprünglich in einem anderen Zusammenhang mit abweichenden Kontextfragen gefällt wurde (vgl. HIGGINS, MCCANN, 1984, S. 26).

   In vielen Fällen haben sich die Probanden jedoch noch keine Meinung zur betreffenden Frage gebildet. Dann ist zunächst die Aufnahme der relevanten Information erforderlich. Hierbei können die Personen jedoch auf Grund des Zeitdrucks nicht alle relevanten Informationen abrufen, so dass nur ein Teil der Information zur weiteren Meinungsbildung herangezogen wird. Welche Information herangezogen wird, kann durch

Kontextfragen bewusst beeinflusst werden. Im Anschluss daran stellt sich für den Probanden die Frage, wie er diese Information verwenden soll.

3. Die *Auswahl* der Antwortkategorie ist der vorletzte Schritt bei der Beantwortung. Dabei kann die Beantwortung sehr stark durch die zur Auswahl vorgegebenen Antwortkategorien beeinflusst werden (vgl. HIPPLER, SCHWARZ, 1987, S. 84).

4. Die *Veränderung* der Antwort auf Grund der Interviewsituation bildet schließlich den Abschluss der kognitiven Prozesse. Dazu zählt z.B. das bekannte Phänomen, dass Versuchspersonen sich selbst als glücklicher einschätzen, wenn der Interviewer eine körperliche Behinderung hat. Eine Erklärung dafür ist, dass der behinderte Interviewer als Vergleichsmaßstab herangezogen wird und somit das eigene Wohlbefinden als besser eingestuft wird.

Auch wenn die Autoren die zur Beantwortung einer Frage nötigen kognitiven Prozesse nur sehr oberflächlich behandeln und vor allem auf die Kontext-Zusammenhänge abzielen, erkennen sie, dass die kognitiven Prozesse von großer Bedeutung sind: „It therefore seems that the understanding of response effects in surveys is best fostered if their cognitive determinants are identified.“ (STRACK, MARTIN, 1987, S. 144)

## 4.5 Der Antwortprozess nach TOURANGEAU et al.

TOURANGEAU, RIPS und RASINSKI stellten bisher das ausführlichste Modell zum Beantwortungsprozess auf. Ihr Modell beruht auf den in Kapitel 4.1 vorgestellten vier Hauptkomponenten:

- Verständnis der Frage
- Abrufen der Information
- Beurteilung der Information
- Beantwortung der Frage

Darüber hinaus gehen die Autoren detailliert auf die Besonderheiten verschiedener Fragestellungen ein. Die erforderlichen Prozesse für das Verständnis der Frage decken sich größtenteils mit den bereits vorgestellten Ansätzen. Daher soll darauf im Folgenden nicht näher eingegangen werden. Interessante Aspekte werden von den Autoren vor allem beim Abrufen und der Beurteilung der Information (Kapitel 4.5.1) und bei der Beantwortung der Frage (Kapitel 4.5.2) beleuchtet.

### 4.5.1 Abrufen und Beurteilung vorhandener Information

Beim Abrufen und der Beurteilung vorhandener Information treten oftmals Interaktionen auf (vgl. TOURANGEAU et al., 2005, S. 10). Das Abrufen von Informationen entspricht in den meisten Fällen der Erinnerung einer Person an bestimmte Sachverhalte. Die Beurteilung der Information hängt wiederum entscheidend davon ab, welche Sachverhalte erinnert werden können.

Die Frage nach den genauen Abläufen bei der Erinnerung bestimmter Ereignisse ist in der aktuellen Forschung nicht einheitlich beantwortet. Auf die unterschiedlichen Modelle zur Funktionsweise des menschlichen Gehirns, wie etwa von TULVING (1983), KOLODNER (1985) oder CONWAY (1996) soll an dieser Stelle nicht eingegangen werden. Einigkeit herrscht jedoch darüber, dass ungenaue oder gar falsche Erinnerungen folgende Ursachen haben können (vgl. TOURANGEAU et al., 2005, S. 82):

- Auskunftspersonen haben die betreffende Information erst gar nicht aufgenommen
- Sie sind nicht bereit, den Erinnerungsprozess zu durchlaufen
- Sie können sich an die betreffende Situation nicht erinnern, sondern nur allgemein an Begebenheiten dieser Art
- Sie können sich nur noch an Teile des Ereignisses erinnern
- Die den Sachverhalt betreffende Information ist im Gedächtnis fehlerhaft gespeichert

Die kognitiven Prozesse sind von der Fragestellung abhängig. In Kapitel 4.5.1.1 sollen zunächst Fragen mit *Zeitbezug* näher betrachtet werden. Andere Strategien verwenden Personen dagegen, wenn nach der *Häufigkeit* von Ereignissen gefragt ist (Kapitel 4.5.1.2). Abschließend soll in Kapitel 4.5.1.3 explizit auf die kognitiven Prozesse bei den in den Wirtschaftswissenschaften besonders bedeutsamen Einstellungsfragen eingegangen werden.

### 4.5.1.1 Fragen mit Zeitbezug

Der Zeitbezug einer Frage ist nur in sehr seltenen Fällen irrelevant. Ausnahmen bilden etwa die Frage nach dem Geschlecht oder Namen der Versuchsperson. In vielen anderen Situationen ist die betreffende Information zeitgebunden, etwa das Alter des Probanden. Zeitgebundene Fragen können sich auf den Zeitpunkt des Interviews beziehen (Interviewzeitpunkt) oder auf der Vergangenheit basieren (Referenzzeitpunkt). Darüber hinaus kann auch die Zeitspanne (Referenzperiode) zwischen Interviewzeitpunkt und Referenzzeitpunkt relevant sein, beispielsweise wenn nach der Häufigkeit von bestimmten Ereignissen innerhalb der Referenzperiode gefragt ist (vgl. TOURANGEAU et al., 2005, S. 63f.). Ein Ereignis könnten beispielsweise die Krankenhausaufenthalte einer Person sein. Diese Begebenheiten selbst können wiederum unterschiedlich lange andauern, weshalb die Zeitdauer einzelner Ereignisse ebenfalls erfasst wird. Abbildung 4.8 veranschaulicht dies nochmals:

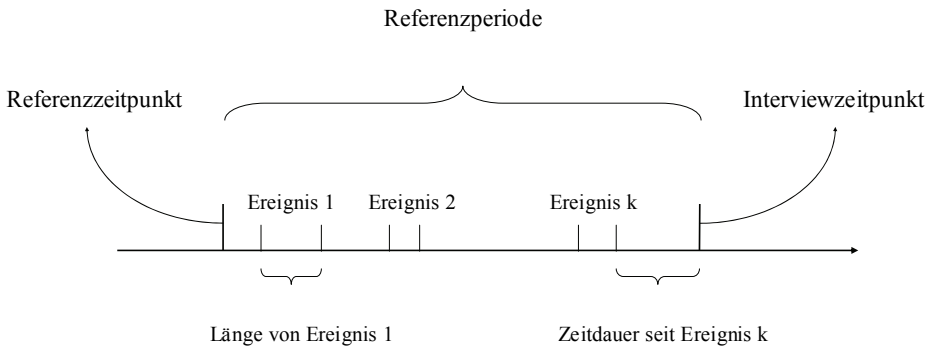


**Abbildung 4.8: Der Zeitbezug einer Frage (in Anlehnung an TOURANGEAU et al., 2005, S. 65)**

Mit Hilfe von Abbildung 4.8 ist eine Einteilung von Fragen mit Zeitbezug in die folgenden Kategorien möglich (vgl. TOURANGEAU et al., 2005, S. 102f.):

- Fragen nach dem Zeitpunkt von bestimmten Ereignissen
- Fragen nach der Zeitdauer von Ereignissen
- Fragen nach der Zeitdauer, die seit dem Ereignis vergangen ist

Bei allen Kategorien können prinzipiell folgende Strategien (S) unterschieden werden (TOURANGEAU et al., 2005, S. 109):

- Abrufen der exakten zeitlichen Information (S1)
- Erinnerung an die zeitliche Abfolge von Ereignissen ohne exaktes Datum (S2)
- Ins Bewusstsein rufen von zusätzlichen Details, aus denen auf zeitliche Aspekte geschlossen werden kann (S3)
- Eindrücke, die auf dem Erinnerungsprozess basieren, aus denen auf zeitliche Aspekte geschlossen werden kann (S4)

Die Probanden versuchen mit allen vier Strategien an Informationen zu gelangen und diese zusammenzuführen. Die einzelnen Wege werden also weder parallel noch alternativ durchlaufen, vielmehr handelt es sich um ineinander greifende Prozesse (vgl. COLLINS, MICHALSKI, 1989, S. 2). Ein Beispiel aus TOURANGEAU, RIPS und RASINSKI (2005, S. 110) verdeutlicht dies:

**Beispiel 4.1**

*Stellt man einer Auskunftsperson die Frage, wann sie ihr Autoradio gekauft hat, so könnte sie folgende Überlegungen anstellen:*

*Zunächst erfolgt der Versuch, sich an das exakte Datum oder die exakte Zeitspanne seit dem Kauf zu erinnern, aber dies ist nicht mehr bekannt (S1).*

*Die Versuchsperson kann sich erinnern, dass sie das Autoradio kurz nach dem Kauf des Autos angeschafft hat (S2). Das Auto ist ein Modell des Jahres 1993, dieses Modell wurde in den Jahren 1992 und 1993 verkauft. Daher kommen nur diese beiden Jahre für den Kauf des Radios in Frage.*

*Darüber hinaus weiß die Person, dass der Autokauf sich etwa drei Monate vor Antritt der aktuellen Arbeitsstelle zutrug (S2).*

*Außerdem weiß sie, dass sie im September 1993 an ihrer neuen Arbeitsstelle angefangen hat (S1).*

*Unter Umständen weiß die Auskunftsperson lediglich, wie lange sie dort schon tätig ist und kann daraus folgern, wann sie die Stelle angetreten hat. Zusätzlich kann sie sich erinnern, dass beim Kauf des Autoradios die Tochter dabei war und sie anschließend mit ihr zum Minigolfspielen ging. Daraus kann sie folgern, dass es im Sommer gewesen sein muss (S3).*

*Damit gelangt die Person insgesamt zur Erkenntnis, dass der Autoradiokauf im Juli oder August 1993 erfolgt sein muss.*

Die in Beispiel 4.1 geschilderten Abläufe sind in Abbildung 4.9 dargestellt. Ereignisse sind in der Abbildung durch Vierecke gekennzeichnet, Daten durch Ellipsen. Gewinnt die Person das Datum auf Grund einer Schlussfolgerung, ist dies durch eine gestrichelte Linie angedeutet. Eine durchgehende Linie signalisiert dagegen, dass sie sich an das exakte Datum erinnert.

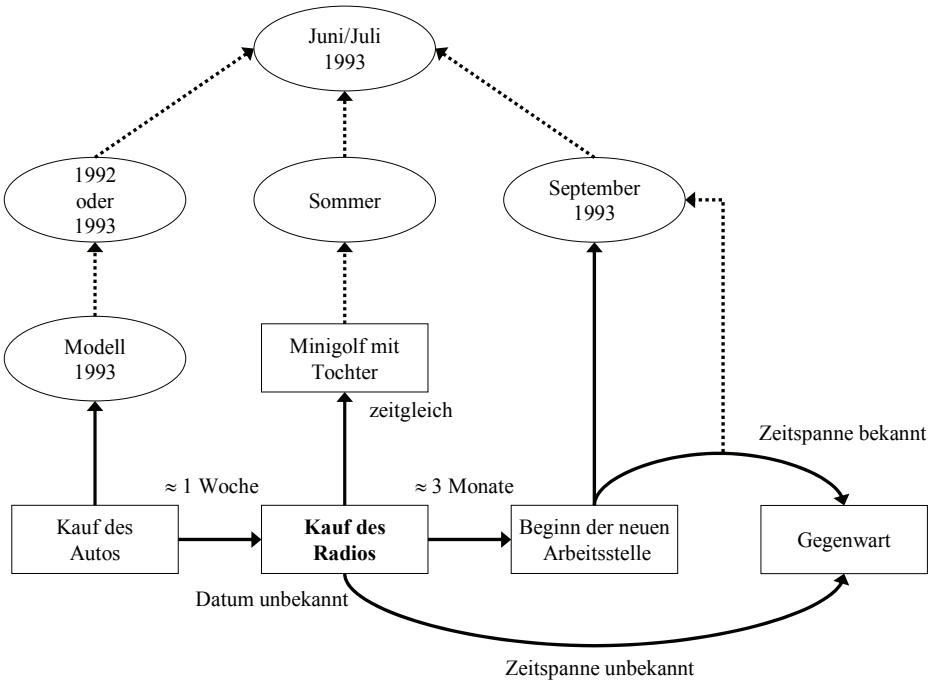


**Abbildung 4.9: Der Prozess des Abrufens von Informationen (in Anlehnung an TOURANGEAU et al., 2005, S. 111)**

Die einzelnen Teilinformationen müssen nicht immer wie in Beispiel 4.1 konsistent sein. Es ist durchaus möglich, dass sich bestimmte Sachverhalte durch unvollständige oder fehlerhafte Erinnerung widersprechen (vgl. TOURANGEAU et al., 2005, S. 111). Darüber hinaus sind die einzelnen Informationen nicht zwangsläufig voneinander abhängig. Es ist demzufolge möglich, die Jahreszeit korrekt wiederzugeben und trotzdem bei der Erinnerung an das Jahr einem Irrtum zu unterliegen (vgl. TOURANGEAU et al., 2005, S. 112).

#### 4.5.1.2 Fragen nach der Häufigkeit von Ereignissen

Bei der Frage nach der Häufigkeit von Ereignissen tritt zusätzlich zu den bereits erörterten Fehlerquellen, dem Gedächtnis und der Einschätzung von Zeitdauern bzw. -punkten, das Problem der numerischen Abschätzung auf. Neben der ursprünglichen Auffassung, dass die Probanden sich an die einzelnen Situationen erinnern und diese dann summieren, wie sie etwa SUDMAN und BRADBURN (1973) vertreten, haben sich inzwischen alternative Möglichkeiten der numerischen Abschätzung in der Literatur etabliert. Diese lassen sich in vier Hauptgruppen einteilen (vgl. TOURANGEAU et al., 2005, S. 146):

- Erinnerung an den exakten Wert (S1)
- Erinnerung an einzelne Situationen (S2)
- Zurückgreifen auf allgemeine Verhaltensmuster (S3)
- Grobe Mutmaßungen (S4)

Wenn dem Probanden der *exakte Wert* ohnehin bekannt ist, kann der Erinnerungsprozess vollständig entfallen. Diese Art der Informationsgewinnung ist demnach insbesondere dann von Bedeutung, wenn die Ereignisse sehr selten oder besonders wichtig sind. So weiß beispielsweise eine Frau die Geburtstage ihrer Kinder auch ohne Erinnerungsprozess (vgl. TOURANGEAU et al., 2005, S. 149).

Bei der Erinnerung an *einzelne Situationen* durchläuft der Proband die gesamten im Kapitel 4.5.1.1 beschriebenen Prozesse und zählt anschließend die Ereignisse innerhalb der Referenzperiode (vgl. NETER, WAKSBERG, 1964, S. 17). Anstatt sich an alle Situationen zusammenhangslos zu erinnern, können Probanden auch

Untergruppen bilden und in diesen die einzelnen Ereignisse zählen (vgl. LESSLER et al., 1989, S. 6). Bei der Frage nach der Anzahl der Restaurantbesuche im letzten Monat würde dies beispielsweise bedeuten, dass der Proband zunächst die Anzahl der Besuche in deutschen Gaststätten bestimmt und anschließend die in ausländischen Restaurants. Insbesondere wenn die Anzahl der Ereignisse sehr groß bzw. die Referenzperiode sehr lang ist, erinnern sich viele Personen nicht an die gesamte Referenzperiode, sondern lediglich an kürzere Zeitabschnitte. Die Anzahl der Ereignisse innerhalb dieses Zeitabschnitts dient dann als Ausgangspunkt für die Hochrechnung auf die gesamte Referenzperiode (vgl. CONRAD et al., 1998, S. 339). Alternativ könnten Probanden auch die seit dem letzten Ereignis vergangene Zeit als Ansatzpunkt verwenden, um auf die Häufigkeit innerhalb der Referenzperiode zu schließen (vgl. LESSLER et al., 1989, S. 6).

Die Erinnerung an einzelne Situationen entfällt, wenn die Ereignisse allgemeiner Natur sind und daher so häufig eintreten, dass die genaue Erinnerung an eine spezielle Information unmöglich ist (vgl. SMITH E.R., 1999, S. 251). In diesen Fällen greifen Versuchspersonen auf *allgemeine Verhaltensmuster* zurück (vgl. TOURANGEAU et al., 2005, S. 148). Eine Befragung zu den Essgewohnheiten der Versuchspersonen zeigt, dass die Ergebnisse für länger zurückliegende Referenzperioden nur unwesentlich schlechter sind (vgl. SMITH A.F., 1991). Dies legt den Schluss nahe, dass sich die Probanden nicht an die exakte Information, sondern an ihre üblichen Essgewohnheiten erinnern (vgl. TOURANGEAU et al., 2005, S. 149). Diese individuelle Information über bestimmte Verhaltensmuster kann ebenfalls durch allgemein bekannte oder von Experten empfohlene Häufigkeiten ersetzt werden, zum Beispiel bei der Frage nach der Anzahl der Impfungen der Kinder (vgl. WILLIS et al., 1999).

Insbesondere bei geschlossenen Fragestellungen können *grobe Mutmaßungen* über die Häufigkeit von Ereignissen die kognitiven Prozesse des Erinnerns ersetzen (vgl. CONRAD et al., 1998, S. 355). In diesem Fall gewinnen die Personen auf Basis der vorgegebenen Kategorien einen *generellen Eindruck* über eine mögliche Antwort. Die mittlere Kategorie dient als Ankerpunkt, den die Probanden als Durchschnittswert aller Personen interpretieren (vgl. SCHWARZ, HIPPLER, 1987, S. 165). In Abhängigkeit des subjektiven Empfindens über die eigene Position im Vergleich zur Allgemeinheit passen die Personen dann ihre Antwort nach oben oder unten an (vgl. TOURANGEAU et al., 2005, S. 150). Ein

Indiz für die Verwendung eines generellen Eindrucks bei offenen Fragestellungen sind überproportional häufig vorkommende runde Zahlenwerte (vgl. CONRAD et al., 1998, S. 358).

Die vier genannten Strategien können, wie mehrere Studien belegen, von den Probanden auch synchron eingesetzt werden (vgl. TOURANGEAU et al., 2005, S. 150). So können durch die Erinnerung an einzelne Situationen gewonnene Informationen durch Berücksichtigung allgemeiner Verhaltensmuster abgeändert werden (vgl. BICKART et al., 1990, S. 198). Da die Exaktheit der gegebenen Antwort sehr stark von der ausgewählten Strategie zur numerischen Abschätzung abhängt, bedarf es einer weiteren Erörterung, welche Einflüsse zur Verwendung bestimmter Strategien führen. In der Literatur sind insgesamt fünf unterschiedliche Aspekte genannt, welche die Wahl der Informationsgenerierung beeinflussen (vgl. BLAIR, BURTON, 1987, S 282):

- Aufwand, den die Erinnerung an spezielle Situationen bereitet
- Motivation
- Zugänglichkeit der Information von einzelnen Ereignissen
- Verfügbarkeit alternativer Möglichkeiten
- Art der Fragestellung

Der kognitive *Aufwand*, den Probanden für die Erinnerung und das Abzählen von Ereignissen aufbringen müssen, steigt deutlich mit der Zahl der tatsächlichen Begebenheiten (vgl. TOURANGEAU et al., 2005, S. 154). Daher greifen nur sehr wenige Personen auf das Abzählen erinnerter Ereignisse zurück, wenn mehr als 10 Situationen reflektiert werden müssen (vgl. BLAIR, BURTON, 1987, S. 285). Die Mehrzahl der Versuchspersonen verwendet jedoch diese Strategie, wenn lediglich ein oder zwei Situationen zu berichten sind (vgl. TOURANGEAU et al., 2005, S. 154). Die Bereitschaft der Probanden, die kognitiven Prozesse zu durchlaufen, kann durch die *Motivation* beeinflusst werden. So steigt der Anteil der Personen, welche diese Strategie verwenden, wenn die Frage als besonders bedeutsam gekennzeichnet ist (vgl. BURTON, BLAIR, 1991, S. 71). Ebenso positiv wirkt es sich aus, wenn die Information im Gehirn für den Probanden leicht abrufbar ist. Diese *Zugänglichkeit* ist erschwert, wenn sich die einzelnen Episoden kaum unterscheiden (BLAIR, BURTON, 1987, S. 282). In diesem Fall

überlagern sich die einzelnen Ereignisse im Gedächtnis, weshalb die Erinnerung an eine konkrete Situation nahezu unmöglich ist (GILLUND, SHIFFRUN, 1984, S. 32). Die Verwendung der übrigen Strategien (exakter Wert, allgemeine Verhaltensmuster, grobe Mutmaßungen) setzt die *Verfügbarkeit* alternativer Informationen voraus. Nur wer den genauen Wert im Gehirn gespeichert hat, kann darauf direkt zurückgreifen (vgl. TOURANGEAU et al., 2005, S. 155). Die *Art der Fragestellung* umfasst abschließend einen breiten Bereich an Möglichkeiten die Informationsgewinnung zu beeinflussen, von der Wortwahl der Frage, der Fragetechnik bis hin zu den einleitenden Kontextfragen (vgl. TOURANGEAU et al., 2005, S. 156).

#### 4.5.1.3 Informationsbeurteilung und Erinnerung bei Einstellungsfragen

Die in den Kapiteln 4.5.1.1 und 4.5.1.2 dargestellten psychologischen Prozesse sind vor allem bei Fragen zu Ereignissen oder bei Fragen zum Verhalten in bestimmten Situationen erforderlich. Sie bilden jedoch auch die Grundlage für die Erklärung von Beantwortungseffekten bei Einstellungsfragen. Bei diesen Fragen ist eine Überprüfung der Exaktheit einer Antwort allerdings unmöglich, da hier keine objektiv nachvollziehbaren Sachverhalte thematisiert sind (vgl. TOURANGEAU et al., 2005, S. 165). Dennoch können bei Einstellungsfragen Phänomene auftreten, die den Schluss nahe legen, dass auch hier Beantwortungseffekte vorliegen. So zeigen beispielsweise zwei im Abstand von einem halben Jahr durchgeführte Studien von ZALLER (1992) zum Thema Abbau des Sozialstaates, dass über die Hälfte der Personen ihre Meinung bei der zweiten Umfrage ändert. Diese Variabilität der Antworten auf Einstellungsfragen ist in vielen Studien belegt (vgl. JUDD, KROSNICK, 1982; JUDD, MILBURN, 1980). Neben der Möglichkeit, dass die Probanden ihre Meinung tatsächlich geändert haben, sind auch Messfehler eine mögliche Ursache für die Veränderung (vgl. TOURANGEAU et al., 2005, S. 169). Daher ist es sinnvoll, die Theorien über die notwendigen Prozesse bei Einstellungsfragen näher zu betrachten.

Neben den ersten Überlegungen zum psychologischen Kontinuum von THURSTONE (1927) ist es vor allem die Arbeit von ALLPORT (1935), die richtungsweisend für erste Modelle zu den kognitiven Prozessen sind (vgl. TOURANGEAU et al., 2005, S. 167). Die Arbeit von ALLPORT (1935) zeigt, dass in der Kindheit gebildete Einstellungen das ganze Leben über unverändert bleiben

können. Daher gehen die ersten Modelle davon aus, dass Einstellungen *existierenden Meinungen* zu bestimmten Sachverhalten entsprechen, die relativ geringen Änderungen unterworfen sind (vgl. TOURANGEAU et al., 2005, S. 167). Sie werden in der Literatur daher auch `File Drawer Models´ genannt (vgl. WILSON, HODGES, 1992, S. 38). File Drawer Modelle entsprechen der Erinnerung an den exakten Wert, wenn nach der Häufigkeit von Ereignissen gefragt wird. Sie gehen davon aus, dass die relevante Information bereits im Gehirn vorliegt (vgl. TOURANGEAU et al., 2005, S. 167). Mit diesen Modellen kann jedoch nicht erklärt werden, weshalb die Antworten bei Einstellungsfragen starken Schwankungen unterworfen sind und warum Kontextfragen einen großen Einfluss nehmen können (vgl. TOURANGEAU et al., 2005, S. 170f.). Neben dieser inzwischen als traditionell bezeichneten Sichtweise haben sich drei weitere Erklärungsansätze gebildet, wie Personen die nötige Information zur Beantwortung von Einstellungsfragen erhalten können (vgl. TOURANGEAU et al., 2005, S. 171 ff). Diese Ansätze sind größtenteils analog zu den bereits betrachteten Überlegungen zur Abschätzung von Häufigkeiten. Man unterscheidet:

- Gefestigte Meinung (S1)
- Situationsbedingte Vorstellungen oder Gefühle zum Sachverhalt (S2)
- Grundlegende Werte (S3)
- Eindrücke und Stereotypen (S4)

Bei manchen Fragestellungen hat eine Person bereits eine *gefestigte Meinung* zu diesem Thema. Diese Antwortstrategie ist vergleichbar zum Abrufen eines exakten Wertes bei Häufigkeitsfragen.

Beruht die Antwort einer Person auf *situationsbedingten Vorstellungen oder Gefühlen* zum Sachverhalt, verdrängt die Einschätzung einzelner Aspekte die Beurteilung der globalen Einstellung. Bei der Frage nach der Zufriedenheit der Person wird etwa die Zufriedenheit im Job oder in der Beziehung bewertet und darauf die generelle Antwort aufgebaut (vgl. MCCLENDON, O´BRIEN, 1988, S. 359). Dieser Prozess ist der Erinnerung an einzelne Situationen bei Häufigkeitsfragen ähnlich (vgl. TOURANGEAU et al., 2005, S. 175).

Ist die Beantwortung auf *grundlegende Wertevorstellungen* zurückzuführen, entspricht dies nahezu der Verwendung von allgemeinen Verhaltensmustern bei Häufigkeitsfragen. Statt die konkrete Fragestellung zu beantworten, greifen die Probanden auf ihre Überzeugungen und Wertevorstellungen zurück (vgl. TOURANGEAU et al., 2005, S. 174). Diese Sichtweise gibt eine mögliche Erklärung für die großen Auswirkungen leichter Veränderungen in der Fragestellung. In einer Umfrage befürworteten beispielsweise deutlich mehr Amerikaner die Unterstützung der Rebellen Nicaraguas, wenn die Fragestellung zusätzlich die Einstellung zur Verhinderung der weiteren Verbreitung des Kommunismus beinhaltet (vgl. ZALLER, 1992, S. 128ff.).

Eine auf *Eindrücken* und *Stereotypen* basierende Antwort ist vergleichbar mit einer groben Mutmaßung bei Häufigkeitsfragen. Anstatt durch den Ablauf der erforderlichen kognitiven Prozesse eine fundierte Meinung abzugeben, beruht die Antwort der Probanden auf Stereotypen oder dem Gefühl, welche Antwort geeignet sein könnte. Diese Strategie wenden Probanden vor allem dann an, wenn sie keine vorgefertigte Meinung abrufen können und Zeit bzw. Motivation fehlt, um eine genaue Beurteilung der zur Verfügung stehenden Informationen vorzunehmen (vgl. SANBONMATSU, FAZIO, 1990, S. 617). Bei der Beurteilung von Politikern im Wahlkampf tendieren beispielsweise viele Personen dazu, die generelle Sympathie der exakten Information vorzuziehen (vgl. LODGE et al., 1989, S. 413). Ein möglicher Grund dafür könnte die Informationsflut im Wahlkampf sein, deren Verarbeitung für die Probanden einen zu großen Aufwand darstellt. Daher verlassen sie sich bei einem Urteil auf ihre Eindrücke oder Stereotypen über das betreffende Befragungsobjekt.

Beispiel 4.2 verdeutlicht nochmals die unterschiedlichen, bei Einstellungsfragen möglichen Informationsgrundlagen.

**Beispiel 4.2**

*Im Rahmen eines Fragebogens werden die Probanden gefragt, ob sie für oder gegen die Legalisierung von Marihuana sind. Die Antworten können auf folgende Weise gebildet werden (vgl. TOURANGEAU et al., 2005, S. 165):*

1. *Die Person greift auf eine bereits **gefestigte Meinung (S1)** zur Legalisierung weicher Drogen zurück.*

2. *Die Person gewinnt ihre Antwort auf Basis **situationsbedingter Vorstellungen (S2)** über die Legalisierung weicher Drogen, etwa der Coffee-Shops in Amsterdam.*

3. *Bei der Fragestellung wird die **grundlegende Wertevorstellung (S3)** der Person zur individuellen Freiheit oder der Fürsorgepflicht des Staates aktiviert. Diese Einstellung bestimmt die Antwort auf die Frage zur Legalisierung von Marihuana.*

4. *Die Person folgt mit ihrer Antwort dem **Eindruck (S4)**, den sie von den Befürwortern bzw. Gegnern der Legalisierung hat.*

Vergleichbar mit der Herleitung der Häufigkeit bestimmter Ereignisse können auch bei Einstellungsfragen alle vier Strategien gleichzeitig zum Einsatz kommen. Die Antwort in Beispiel 4.2 kann demnach alle genannten Aspekte berücksichtigen. Als Basis für die Beantwortung kommen nur für den Probanden verfügbare Informationen in Frage (vgl. TOURANGEAU et al., 2005, S. 179). Eine Person kann eine gefestigte Meinung nur dann ausdrücken, wenn sie diese bereits gebildet hat. Neben der Verfügbarkeit nimmt die Motivation entscheidenden Einfluss auf die Art der Informationsbeschaffung, da der kognitive Aufwand nicht bei allen Wegen identisch ist (vgl. TOURANGEAU et al., 2005, S. 179). Aus diesen unterschiedlichen Möglichkeiten resultieren erhebliche zeitliche Inkonsistenzen der Antworten. Häufig reicht eine leichte Abänderung der Frageform oder eine zusätzliche einleitende Kontextfrage aus, um einen anderen Prozess der Informationsbeurteilung des Probanden zu aktivieren (vgl. TOURANGEAU et al., 2005, S. 171 ff.).

Das sogenannte *Belief-Sampling-Modell* geht davon aus, dass die Einstellung einer Person zu einem Sachverhalt eine Art Datenbasis – bestehend aus gefestigten Meinungen (S1), situationsbedingten Gefühlen und Eindrücken (S2), grundlegenden Wertevorstellungen (S3) und Eindrücken und Stereotypen (S4) – bildet (vgl. TOURANGEAU et al., 2005, S. 179). Welche dieser Größen in die konkrete Fragebeantwortung einfließen, hängt von der aktuellen Verfügbarkeit ab. Diese ist durch den Wortlaut der Frage, die dem Probanden gegebenen Anweisungen (vgl. OTTATI et al., 1989, S. 404) und durch Kontextfragen (vgl. TOURANGEAU, RASINSKI, 1988, S. 299) beeinflusst. Eine bereits gefestigte Meinung wird von den Probanden insbesondere dann zur Beantwortung herangezogen, wenn sie sehr stark internalisiert ist (vgl. FAZIO, 1989, S. 153). In

Abbildung 4.10 ist dies bei Fragestellung 1 der Fall, die ausschließlich durch die Aktivierung der gefestigten Meinung beantwortet wird. Falls keine gefestigte Meinung vorliegt, entsteht ein erstes Urteil auf Basis einer der anderen Strategien, welche im Anschluss daran durch weitere Informationen bestätigt oder verworfen wird (vgl. TOURANGEAU et al., 2005, S. 180). In Abbildung 4.10 zeigt Fragestellung 3 diese Urteilsfindung. Fragestellung 2 zeigt eine weitere Möglichkeit. Die Person greift auf ihre gefestigte Meinung zu einer Frage zurück, berücksichtigt aber zusätzlich noch weitere Informationen.

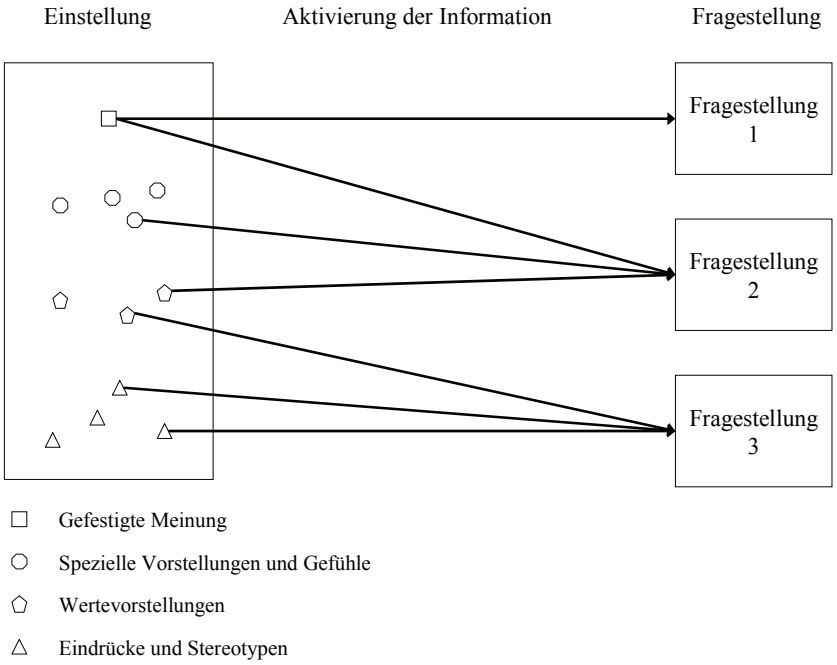


**Abbildung 4.10: Die abgerufene Information beim Belief-Sampling-Modell**

Die zeitliche Konsistenz der Beantwortung hängt beim Belief-Sampling-Modell von folgenden Größen ab:

- Zeitliche Konsistenz bei der Bewertung identischer Aspekte: Eine Einstellungsfrage wird demnach nur dann konsistent beantwortet, wenn auch die einzelnen Aspekte konsistent beurteilt werden. Bei der Frage nach der Legalisierung von Marihuana könnte dies etwa die Konsistenz bei der Bewertung der Coffee-Shops in Amsterdam sein.

- Der Korrelation der im Gedächtnis verankerten Meinungen (S1), Gefühle (S2), Wertevorstellungen (S3) und Stereotypen (S4): Liefern diese alle die gleichen Ergebnisse, ist die gemessene Einstellung unabhängig davon, welche Informationsstrategie die Person gewählt hat.

- Dem Anteil der Aspekte, die bei einer zweiten Messung erneut aktiviert werden: Wenn bei einer späteren Befragung erneut die gleichen Aspekte die Informationsgrundlage bilden, so ist lediglich die zeitliche Konsistenz bei der Bewertung von Bedeutung.

Das Belief-Sampling-Modell berücksichtigt die Möglichkeit, dass Einstellungen auf Grund gefestigter Meinungen gebildet werden, und stellt damit keinen Widerspruch zu den Arbeiten von ALLPORT (1935) dar. Die Beurteilung kann jedoch zusätzlich von äußeren Einflüssen abhängen, da diese die Verfügbarkeit der Information und damit die in den Beurteilungsprozess eingehenden Aspekte beeinflussen. Somit liefert das Belief-Sampling-Modell eine Erklärung für die empirisch beobachtbare zeitliche Inkonsistenz von Antworten.

### 4.5.2 Beantwortung der Frage

Beim letzten Schritt des Antwortprozesses erfolgt die Beantwortung der Frage, welche meistens auf der abgerufenen und bewerteten Information beruht. Es kann jedoch auch zu einer bewussten Anpassung der Antwort durch den Probanden kommen. Insbesondere bei geschlossenen Fragestellungen taucht weiterhin das Problem auf, die Antwort in die vorgegebenen Skalen oder Kategorien abzubilden (vgl. TOURANGEAU et al., 2005, S. 13). In Abbildung 4.11 sind diese beiden zentralen Bestandteile der Beantwortung dargestellt.

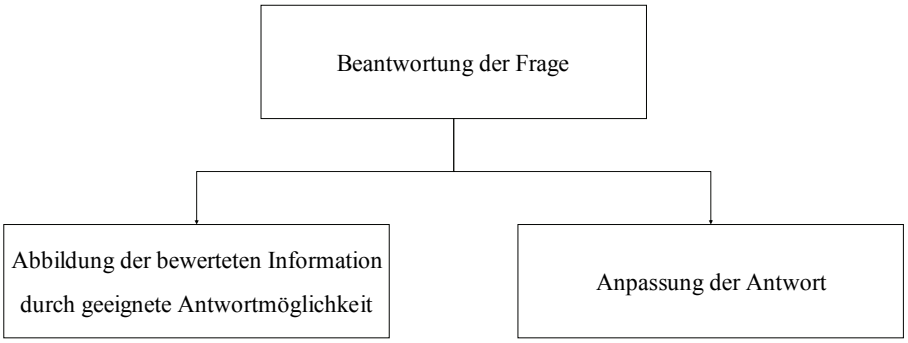


**Abbildung 4.11: Probleme bei der Beantwortung der Frage**

Die Anpassung der Antwort ist bereits ausführlich beim Ansatz CANNELL, MILLER und OKSENBERG (1981) dargestellt (S. 105 ff. in dieser Arbeit). In Kapitel 4.1 wurde die Abbildung der bewerteten Information in eine geeignete Antwortmöglichkeit bislang nur kurz angesprochen. An dieser Stelle soll nun das konkrete Modell von TOURANGEAU et al. (2005) ausführlich erläutert werden.

Die Formulierung der Antwort und die damit für die Versuchsperson verbundene kognitive Herausforderung ist von der Fragestellung abhängig. Es müssen folgende Fragestellungen unterschieden werden:

- Geschlossene Fragestellung *ohne* Rangfolge der Antwortkategorien (Kapitel 4.5.2.1)
- Geschlossene Fragestellungen *mit* Rangfolge der Antwortkategorien (Kapitel 4.5.2.2)
- *Offene* Fragestellungen (Kapitel 4.5.2.3)

### 4.5.2.1 Antwortkategorien ohne Rangfolge

Bei der Studie von KROSNICK und ALWIN (1987) wurde bereits kurz der Reihenfolgeeffekt erläutert. Der Reihenfolgeeffekt besagt, dass Personen innerhalb einer Liste von Antwortmöglichkeiten die Kategorien am Anfang der Liste bevorzugt auswählen. Die dafür verantwortlichen psychologischen Prozesse sind jedoch kaum überprüfbar (vgl. TOURANGEAU et al., 2005, S. 253). Neben der ursprünglichen Erklärung von KROSNICK und ALWIN (1987), dass hierfür die Motivation und die intellektuellen Fähigkeiten der Versuchspersonen

entscheidend sind, zeigen SCHWARZ, HIPPLER und NOELLE-NEUMANN (1991) einen weiteren Ansatz zur Beschreibung dieses Effekts auf. Eine Antwortkategorie liefert demzufolge zusätzliche Argumente, welche den Probanden beeinflussen (vgl. SCHWARZ et al., 1991, S. 189). Da die zuerst aufgelisteten Antwortkategorien schon zur Meinungsbildung führen können, werden die zum Schluss genannten Antwortkategorien nicht mehr in Betracht gezogen (vgl. TOURANGEAU et al., 2005, S. 252). Dieser Reihenfolgeeffekt tritt selbst dann auf, wenn lediglich zwei Kategorien zur Auswahl stehen (vgl. SCHWARZ et al., 1991, S. 191).

Im Rahmen einer mündlichen Befragung ist ebenfalls ein Reihenfolgeeffekt zu beobachten. Allerdings führt er in diesem Fall zu einer häufigeren Auswahl der zuletzt genannten Antworten. Eine mögliche Erklärung dafür liefern KROSNICK und ALWIN (1987, S. 203): Bei dieser Form der Befragung können die Probanden die Antworten nicht direkt bewerten, da sie zu schnell vorgelesen werden. Die zuletzt genannten Kategorien bleiben jedoch im Gedächtnis, weshalb die Versuchspersonen bei der Beantwortung auf diese zurückgreifen.

Eine weitere Problematik entsteht bereits durch die reine Vorgabe der Kategorien. Untersuchungen belegen, dass bei bestimmten Fragestellungen der Anteil der Personen, die ihre Meinung zum Sachverhalt äußern, von 42% auf 4% sinkt, wenn die Antwortkategorie ´weiß nicht´ vorhanden ist (vgl. BISHOP et al., 1986, S. 246). Es besteht also die Tendenz, dass Probanden eine Antwortmöglichkeit auswählen, obwohl sie nur in Ansätzen ihrer Meinung entspricht, da die angegebenen Kategorien als einzig mögliche Alternativen betrachtet werden (vgl. SEARLE, 1969, S. 126). Daher gründet die Beantwortung nicht ausschließlich auf der Informationsbeurteilung, sondern hängt sehr stark von den vorgegebenen Kategorien ab. Dieses Phänomen betrifft somit das ursprüngliche Frageverständnis.

#### 4.5.2.2 Antwortkategorien mit Rangfolge

In der Literatur am ausführlichsten diskutiert ist das Antwortverhalten, wenn die Kategorien eine natürliche Rangfolge beinhalten, etwa durch verbale Beschreibungen von ´immer´ bis ´nie´ oder durch konkrete numerische Angaben. Die ersten, auf PARDUCCI (1965, 1974) zurückgehenden Überlegungen zeigen, dass die Probanden die zu bewertenden Objekte so anordnen, dass die extremen

Stimuli in die extremen Kategorien fallen (vgl. PARDUCCI, 1974, S. 129). Anschließend werden die Objekte in die einzelnen Kategorien so eingeteilt, dass alle Kategorien etwa gleich besetzt sind (vgl. PARDUCCI, 1974, S. 130). Demzufolge müssten eigentlich die Häufigkeiten in den einzelnen Kategorien gleichverteilt sein. Dieses Prinzip gilt auch, wenn lediglich ein Stimulus zu bewerten ist, beispielsweise die Arbeit des Außenministers. In diesem Fall dient dem Probanden der gedankliche Vergleich mit weiteren Politikern oder ehemaligen Außenministern als Anhaltspunkt für die Einschätzung des zu beurteilenden Objekts (vgl. TOURANGEAU et al., 2005, S. 240).

Obwohl die Grundlagen dieses Modells empirisch belegt sind (vgl. etwa DAAMEN, BIE, 1992, S. 111), können damit nicht alle vorkommenden Phänomene erklärt werden. So zeigen mehrere Untersuchungen, dass bei der Bewertung von Personen überwiegend positive Einschätzungen bezogen auf deren Charaktereigenschaften auftreten (vgl. LANDY, FARR, 1980, S. 78ff. oder SEARS, 1983, S. 240). Welche psychologischen Prozesse für diese Verzerrung verantwortlich sind, ist in der Literatur jedoch uneinheitlich beantwortet. Ein Teil der Autoren folgert daraus, dass Personen prinzipiell positive Einstellungen gegenüber Mitmenschen haben, so lange keine gegenteilige Information vorhanden ist (vgl. ZAJONC, 1968, S. 361). Andere Autoren gehen davon aus, dass Personen deshalb mit negativen Einschätzungen zurückhaltend sind, weil sie nicht als besonders negativ oder kritisch erscheinen wollen (vgl. TOURANGEAU et al., 2005, S. 241).

Die überproportionale Verwendung der mittleren Kategorien und die Vermeidung von Extremantworten widerspricht ebenfalls dem ursprünglichen Modell von PADUCCI (1974) (vgl. TOURANGEAU et al., 2005, S. 244). Dieser Umstand resultiert aus der Neigung der Auskunftspersonen, besonders positive Objekte schlechter zu bewerten bzw. negative Stimuli besser zu bewerten (vgl. POULTON, 1989, S. 23).

Ein weiterer sehr interessanter Effekt, der eine Gleichverteilung verhindert, entsteht, wenn die verbal beschriebenen Antwortkategorien zusätzlich mit Zahlen beschriftet sind (vgl SCHWARZ et al., 1991, S. 190). Die bereits erläuterte Tendenz, positive Kategorien anzukreuzen, wird auch von dieser Studie unterstützt (vgl. TOURANGEAU et al., 2005, S. 241). Variiert man allerdings die ursprüngliche Zahlenskala von 0 bis 10 auf eine Skala von -5 bis +5 bei gleicher verbaler Beschreibung, so ist die Verschiebung zu den positiven Kategorien besonders gravierend (vgl. SCHWARZ et al., 1991, S. 190). Die Autoren führen dies auf die

unterschiedliche Bedeutung zurück, die Probanden einer Skala mit negativem Beginn im Gegensatz zu einer Skala, die mit Null beginnt, zumessen (vgl. SCHWARZ et al., 1991, S. 191). Beginnt die negativste Kategorie mit einer Null, so liegt der Schluss nahe, dass es sich hierbei lediglich um das logische Gegenstück zur positivsten Kategorie handelt. Fragt man beispielsweise nach dem Erfolg im Leben, so suggeriert die Null lediglich, dass die betreffende Person keinen Erfolg hatte. Ein negativer Zahlenwert impliziert dagegen, dass die negativste Kategorie das genaue Gegenteil der positiven beschreibt, also dass eine Person im Leben völlig gescheitert ist (vgl. TOURANGEAU et al., 2005, S. 243). Daher sind in diesem Fall die Antworten stärker in Richtung des positiven Endes der Skala verschoben. Dieser Effekt wird etwas abgeschwächt, wenn zusätzlich die mittlere Antwortkategorie verbal beschriftet ist (vgl. TOURANGEAU et al., 2005, S. 243).

Wie sehr die Zuordnung von Zahlenwerten zu Antwortkategorien die Versuchspersonen beeinflusst, zeigt auch das folgende Beispiel bei dem gefragt wird, wie viel Zeit eine durchschnittliche Person vor dem Fernseher verbringt. In Abbildung 4.12 ist die Häufigkeitsverteilung für diese Fragestellung bei unterschiedlichen Beschriftungen der Antwortkategorien angegeben (vgl. SCHWARZ et al., 1985, S. 391).

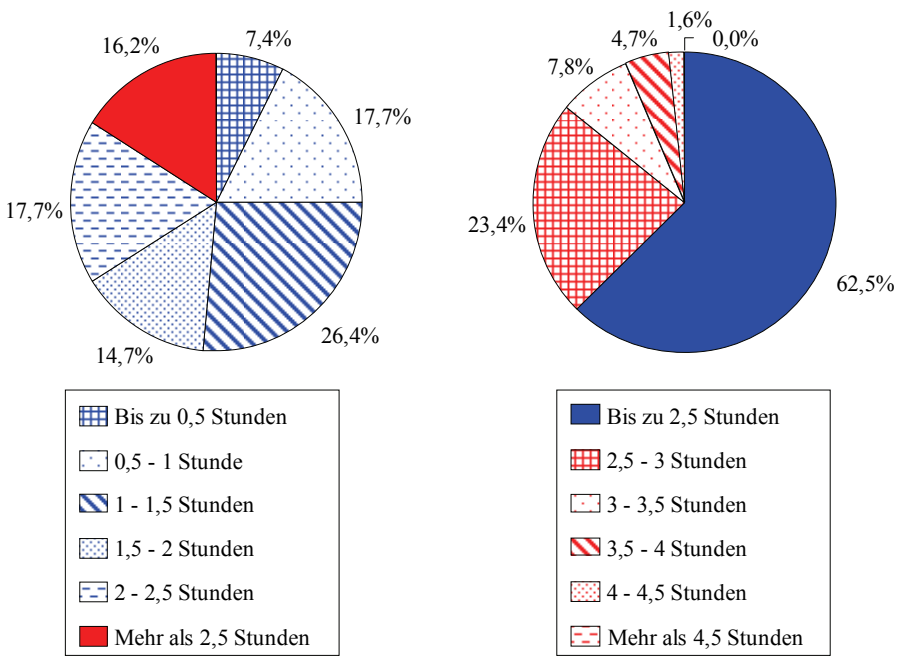


**Abbildung 4.12: Abhängigkeit der Antworten von der Kategorienbeschriftung**

Der Anteil der Probanden, die angeben, länger als 2,5 Stunden pro Tag fernzusehen, wird durch eine Veränderung der Fragestellung mehr als verdoppelt (von 16,2% auf 37,5%). Die Gründe hierfür sind bereits ausführlich diskutiert. So können der Versuchsperson die zur Erinnerung der exakten Zeitdauer notwendigen Prozesse als zu aufwendig erscheinen, weshalb die tatsächliche Fernsehdauer als Informationsgrundlage nicht in Frage kommt. Stattdessen zieht die Person einen Vergleichsmaßstab heran, beispielsweise die vermutete Position innerhalb der Population und formuliert darauf aufbauend die Antwort. Sie entscheidet sich demzufolge unabhängig von der numerischen Beschriftung der Kategorie, lediglich auf Basis der subjektiven Einschätzung der restlichen Bevölkerung. Die Vermeidung der Extremantworten ist eine weitere Ursache für die großen Unterschiede in den Ergebnissen der beiden Fragestellungen.

Die Anzahl der Antwortkategorien hat ebenfalls Einfluss auf die mit dem Item einhergehenden Schwierigkeiten für die Probanden. Während eine zu geringe Anzahl an Kategorien den Personen eine differenzierte Meinungsäußerung unmöglich macht, werden die Probanden bei einer zu großen Kategorienzahl intellektuell überfordert (vgl. TOURANGEAU et al., 2005, S. 249). Wie viele Kategorien im Einzelfall sinnvoll sind, muss noch genauer erörtert werden, da hier in der Literatur unterschiedliche Auffassungen vertreten werden (siehe Kapitel 5).

Diese Probleme bei geschlossenen Fragen lassen es sinnvoll erscheinen, bei der Fragebogenkonstruktion vor allem offene Fragestellungen zu verwenden. Jedoch sind auch bei dieser Form der Befragung Störeinflüsse zu beobachten.

### 4.5.2.3 Offene Fragestellungen

Offene Fragestellungen sind insbesondere dann geschlossenen vorzuziehen, wenn eine präzise Antwort leicht gegeben werden kann oder sehr viele unterschiedliche Antwortalternativen bestehen (vgl. TOURANGEAU et al., 2005, S. 232). Dies gilt im Besonderen, wenn eine konkrete Zahl gefragt ist. Ein typisches Beispiel ist das Alter, denn hier würden Antwortkategorien für den Probanden keine Erleichterung darstellen, sondern lediglich aufgrund der Zusammenfassung verschiedener Altersstufen zu Informationsverlust führen. Betrachtet man jedoch die durch eine offene Fragestellung generierten Antworten genauer, so erkennt man, dass auch diese nicht gleichverteilt sind. Bei der Frage nach dem Alter

treten beispielsweise alle durch fünf teilbaren Zahlen gehäuft auf (vgl. SHRYOCK et al., 1976, S. 115). Besonders auffällig ist dieses Phänomen bei der Frage nach der Anzahl der Geschlechtspartner. Eine runde Zahl verwenden insbesondere jene Personen, die von sehr vielen Partnern berichten (vgl. TOURANGEAU et al., 1997, S. 209). Unter den Probanden, die mehr als 10 Geschlechtspartner angeben, nennen beispielsweise 62,4% ein Vielfaches von fünf (vgl. TOURANGEAU et al., 2005, S. 234). Dieser Effekt mag durch die intime Fragestellung entstehen. Die Probanden beantworten die Frage absichtlich mit gerundeten Werten, um zu signalisieren, dass es sich dabei nur um eine gerundete Zahl handelt. Allerdings ist dieser Rundungseffekt auch bei trivialen Fragen, wie dem zeitlichen Abstand zwischen zwei Interviews, zu beobachten. In diesem Fall antworten überproportional viele Personen mit sieben oder dreißig Tagen (vgl. HUTTENLOCHER et al., 1990, S. 201). Als Hauptursache für die Verwendung gerundeter Ergebnisse gilt daher, dass die Erinnerung an den exakten Wert sehr aufwendig oder unmöglich ist (vgl. HUTTENLOCHER et al., 1990, S. 199). Deshalb ist die Verwendung gerundeter Werte vor allem auf die unpräzise Information zurückzuführen und nicht auf Probleme mit der geeigneten Abbildung der Information in eine Antwort (vgl. TOURANGEAU et al., 2005, S. 235).

Obwohl die gerundeten Ergebnisse wegen Problemen bei der Informationsbeschaffung bzw. -beurteilung entstehen, muss in diesem Kapitel näher darauf eingegangen werden. Dies liegt an der ungleichen Verteilung der runden Werte (vgl. TOURANGEAU et al., 2005, S. 238). Wie bereits angesprochen greifen Personen insbesondere dann auf gerundete Werte zurück, wenn sie sehr viele Ereignisse berichten müssen. Die daraus resultierende Gefahr der Verzerrung verdeutlicht Beispiel 4.3:

**Beispiel 4.3**

*Bei der Frage nach der Anzahl der Geschlechtspartner ist folgender Zusammenhang zu beobachten (vgl. TOURANGEAU et al., 1997, S. 209): Bei bis zu neun Partnern ist kein Anzeichen für die Verwendung von Rundungen zu erkennen. Gehäuft treten vor allem die Zahlen 10, 15, 20, 30 und 50 auf. Im Folgenden soll deshalb unterstellt werden, dass alle Probanden eine exakte zweistellige Zahl auf einen dieser Werte runden. Daher entspricht jede dieser Zahlen einem Intervall, welches nachfolgend angegeben ist:*

*10* ⇒ *[10 ; 12]* *15* ⇒ *[13 ; 17]*

*20* ⇒ *[18 ; 24]* *30* ⇒ *[25 ; 39]*

*50* ⇒ *[40 ; ∞)*

Wie das Beispiel verdeutlicht, nehmen die Intervallbreiten stetig zu. Daher werden tendenziell mehr Personen ihre Antwort abrunden, selbst wenn alle subjektiv exakt runden. Ein auf Basis dieser Werte gebildeter Mittelwert wird den tatsächlichen daher unterschätzen. Darüber hinaus können Probanden auch von den mathematischen Rundungsregeln abweichen und systematisch auf- oder abrunden (vgl. HUTTENLOCHER, 1990, S. 205).

## 4.6 Zusammenfassung und Einordnung

Wie die in diesem Kapitel vorgestellten Studien zeigen, sind viele unterschiedliche psychologische Prozesse für den Probanden erforderlich, um zu einer adäquaten Fragebeantwortung zu gelangen. Diese Vielzahl psychologischer Prozesse erklärt, warum Beantwortungseffekte bei einer Messung entstehen.

Deshalb muss bei der konkreten Ausgestaltung einer Ratingskala berücksichtigt werden, welchen Einfluss sie auf

- das Verständnis der Frage,
- das Abrufen und Beurteilen relevanter Information und
- die Beantwortung der Frage hat.

Für das Verständnis der Frage kann die verbale bzw. numerische Beschriftung der Antwortkategorien problematisch sein. Das Abrufen und die Beurteilung relevanter Information werden beeinflusst, wenn positive und negative Antwortkategorien asymmetrisch sind. Daraus könnten die Probanden Rückschlüsse auf die erwartete Antwort ziehen. Das größte Problem stellt allerdings die Zahl der Antwortkategorien in Hinsicht auf die Beantwortung der Frage dar. Dabei muss Folgendes berücksichtigt werden:

- Ist die Anzahl zu gering gewählt, kann der Proband seine bewertete Information nicht geeignet in eine Antwortkategorie abbilden.

- Bietet man dagegen der Auskunftsperson zu viele Antwortkategorien an, führt dies zu unterschiedlichen Abbildungen der einzelnen Personen.

In Kapitel 5 wird der aktuelle Stand der Forschung zur Gestaltung von Ratingskalen im Folgenden explizit wiedergegeben.

# Kapitel 5

# Forschungsergebnisse zur Gestaltung von Ratingskalen

In Kapitel 3 dieser Arbeit sind verschiedene Möglichkeiten zur Gestaltung der Messung im wirtschaftswissenschaftlichen Kontext vorgestellt worden. Ein zentrales Messinstrument ist dabei die Ratingskala. Ratingskalen bilden das Fundament für die meisten komplexen Skalierungsverfahren. Viele der in Kapitel 3 behandelten ein- und mehrdimensionalen Verfahren greifen auf die Ratingskala zurück, zum Beispiel Semantisches Differential, Likert-Skala, Modell von FISHBEIN, Modell von TROMMSDORFF, MDS und Strukturgleichungsmodelle.

Die Frage, wie eine Ratingskala konkret auszugestalten ist, beschäftigt die Literatur bereits über mehrere Jahrzehnte (vgl. KROSNICK, FABRIGAR, 1997, S. 141). Dabei wurden sehr viele Fragen zur Ausgestaltung einer Ratingskala untersucht, insbesondere die folgenden drei Teilbereiche sind dabei von Bedeutung:

- Beschriftung der Antwortkategorien (Kapitel 5.1)
- Verwendung der Antwortkategorie ´Keine Meinung´ (Kapitel 5.2)
- Anzahl der Antwortkategorien (Kapitel 5.3)

Die Bewertung verschiedener Ratingskalen basiert bei nahezu allen Autoren auf den in Kapitel 3.4 (S. 73-95) erläuterten Kriterien zur Beurteilung der Qualität von Messergebnissen. Besonders häufig werden die Validität und Reliabilität der Messergebnisse als Maßstab herangezogen.

## 5.1 Beschriftung der Antwortkategorien

Bei der Beschriftung der Antwortkategorien stehen prinzipiell zwei Konzepte zur Verfügung: Die Beschriftung mit Zahlen oder die verbale Bezeichnung der Antwortkategorien (vgl. HAMMANN, ERICHSON, 2000, S. 274). Eine Mischform dieser beiden Konzepte liegt vor, wenn alle Antwortkategorien numerisch beschriftet sind und die extremen Antwortkategorien zusätzlich verbal beschrieben werden (vgl. KROSNICK, FABRIGAR, 1997, S. 149).

Bei der verbalen Bezeichnung der Antwortkategorien ist darauf zu achten, dass diese vom Probanden als numerisch äquidistant aufgefasst werden (vgl. GIERL, 1995, S. 45). Wenn dies gelingt, ist die Behandlung der Ratingskala als intervallskaliertes Merkmal zulässig (vgl. HAMMANN, ERICHSON, 2000, S. 274). Damit wäre es zu rechtfertigen, dass lediglich die extremen Antwortkategorien verbal zu beschriften sind, da die präzise äquidistante Beschreibung aller Kategorien numerisch einfacher umgesetzt werden kann (vgl. KROSNICK, FABRIGAR, 1997, S. 149). Außerdem sind die kognitiven Ansprüche an den Probanden bei einer verbalen Beschriftung aller Kategorien höher und können insbesondere bei mündlichen Befragungen zu einer Überforderung der Probanden führen (KROSNICK, FABRIGAR, 1997, S. 149).

Andererseits drücken Menschen im Alltagsleben ihre Gefühle und Einstellungen nicht durch Zahlen aus, sondern greifen auf verbale Umschreibungen zurück (vgl. KROSNICK, FABRIGAR, 1997, S. 149). Daher kann die verbale Beschreibung aller Antwortkategorien sinnvoll sein. Darüber hinaus weisen die einzelnen Antwortkategorien keine eigenständige Bedeutung auf, wenn sie lediglich numerisch beschriftet sind. Mit einer verbalen Beschriftung der Antwortkategorien kann dieses Problem gelöst werden und dadurch unter Umständen eine höhere Güte der Antworten erreicht werden.

Unter Abwägung dieser Aspekte erscheint die verbale Beschriftung aller Antwortkategorien sinnvoll. Diese Ansicht stützt sich auf folgende vier Gesichtspunkte:

- Die meisten Studien zu diesem Thema kommen zum Ergebnis, dass die verbale Beschreibung aller Antwortkategorien die *Reliabilität* erhöht (vgl. ALWIN, KROSNICK, 1991). Insbesondere bei Auskunftspersonen mit

niedriger Bildung steigt die Reliabilität deutlich an, wenn alle Antwortkategorien verbal beschriftet sind (vgl. KROSNICK, BERENT, 1993). Lediglich einige wenige Studien kommen zu dem Schluss, dass die Reliabilität durch verbale Beschriftungen nicht beeinflusst werden kann (vgl. FINN, 1972).

- Nahezu alle Studien gelangen zum Resultat, dass die *Validität* durch verbale Beschriftungen zunimmt (vgl. KROSNICK, BERENT, 1993; ANDREWS, 1984).

- Ein weiterer positiver Effekt der verbalen Bezeichnung ist die geringere Auswirkung der in Kapitel 4.4 (S. 109-111) diskutierten Kontexteffekte. Die Studie von WEDELL (1990) zeigt, dass diese deutlich geringeren Einfluss auf die Befragungsergebnisse besitzen, wenn verbale Beschriftungen vorgenommen werden.

- Darüber hinaus existieren Studien, die direkt danach fragen, welche Formulierung die Versuchspersonen präferieren. Die Mehrzahl der Probanden gibt dabei an, die verbale Ausformulierung zu bevorzugen (vgl. WALLSTEN et al., 1993).

## 5.2 Verwendung der Antwortkategorie ´Keine Meinung´

Ob bei einer Ratingskala die Verwendung der Antwortkategorie ´Keine Meinung´ vorteilhaft ist, wird in der Literatur seit langer Zeit diskutiert. Gestützt wird diese Ansicht dadurch, dass bei mehrmaligem Wiederholen von Einstellungsfragen häufig Meinungsänderungen der Probanden zu beobachten sind. Eine mögliche Erklärung dafür könnte sein, dass Personen Fragen selbst dann beantworten, wenn sie zum betreffenden Thema gar keine Meinung haben. Diese Sichtweise ist auch durch weitere Studien belegt (vgl. BISHOP et al., 1986).

Der Schluss liegt nahe, dass die Probanden die Fragen nur deshalb beantworten, weil sie durch den Fragebogen dazu aufgefordert werden: „…[Respondents] play by the rules of the game“ (SCHUMAN, PRESSER, 1981, S. 299). Es erscheint somit notwendig, den Auskunftspersonen die Möglichkeit einzuräumen, zu einem

bestimmten Thema keine Meinung zu äußern. Durch die zusätzliche Angabe der Kategorie ´Keine Meinung` oder ´Weiß nicht´ erhöht sich nachweislich die Zahl der Personen, die zugeben, zu diesem Thema keine Meinung abgeben zu können (vgl. BISHOP et al., 1986).

Die Angabe der Kategorie ´Keine Meinung´ führt allerdings auch zu negativen Effekten. Bei einer ungeraden Anzahl der Antwortkategorien ist es für Personen möglich, Indifferenz auszudrücken, indem sie die mittlere Kategorie ankreuzen. Bei geradzahligen Antwortmöglichkeiten ist dies jedoch unmöglich. In diesem Fall neigen Probanden bei Indifferenz zur Antwort ´Keine Meinung´, was aber nicht ihrer tatsächlichen Überzeugung entsprechen muss (vgl. KROSNICK, FABRIGAR, 1997, S. 154). Ein weiterer Grund für die Probanden, die Kategorie ´Keine Meinung´ anzukreuzen, lässt sich auf das in Kapitel 4.3 vorgestellte ´Satisficing-Prinzip´ zurückführen (S. 108 f.). Die Auskunftspersonen wählen demnach diese Kategorie aus Gründen der Einfachheit, weil ihnen zur aufrichtigen Beantwortung der Frage die Motivation oder Fähigkeit fehlt (vgl. KROSNICK, 1991).

Für den Ausschluss der Antwortkategorie ´Keine Meinung´ sprechen vor allem zwei Gründe:

- Ergebnisse von Studien legen die Vermutung nahe, dass durch die Angabe der Kategorie ´Keine Meinung´ die Reliabilität der Skalen allenfalls geringfügig gesteigert werden kann (vgl. KROSNICK, BERENT, 1990).

- Es ist davon auszugehen, dass die Motivation der Probanden in empirischen Studien nicht besonders hoch ist. Die Gefahr erscheint daher sehr groß, dass die Versuchspersonen die Kategorie ´Keine Meinung´ aus Gründen der Einfachheit wählen, weil die aufrichtige Beantwortung ihnen zu aufwendig erscheint.

Einen weiteren Ansatz, die Kategorie ´Keine Meinung´ zu umgehen, liefern KROSNICK und FABRIGAR (1997). Statt diese Antwortkategorie anzugeben, können die Personen direkt gefragt werden, welche Bedeutung der befragte Sachverhalt für sie hat (vgl. KROSNICK, FABRIGAR, 1997, S. 158). Auf diese Weise kann man ebenfalls einschätzen, ob ein Proband die Frage rein zufällig beantwortet hat, ohne die Nachteile der Kategorie ´Keine Meinung´ in Kauf nehmen zu müssen.

## 5.3 Anzahl der Antwortkategorien

Die in der Literatur am ausführlichsten diskutierte Frage bezüglich einer Ratingskala ist die optimale Anzahl der Antwortkategorien. Eine grundlegende Überlegung ist dabei, ob eine ungerade Anzahl Antwortkategorien verwendet werden soll, um damit den Versuchspersonen eine mittlere Kategorie anzubieten. Diese mittlere Antwortkategorie bietet den Versuchspersonen die Möglichkeit, eine neutrale Haltung zu einem Thema adäquat zu formulieren. Fehlt diese Alternative, sind die Probanden unter Umständen gezwungen, eine Meinung abzugeben, die nicht ihrer persönlichen Einstellung entspricht, oder die Frage überhaupt nicht zu beantworten. Es erscheint daher auf den ersten Blick vorteilhaft, eine ungerade Anzahl zu verwenden. Diese Ansicht vertritt eine Vielzahl von Autoren, die bei ihren Studien ausschließlich mit einer ungeraden Anzahl Antwortkategorien arbeiten. Eine Literaturrecherche von COX (1980, S. 407-422) zu diesem Thema zeigt, dass bis dato fast alle Autoren von einer ungeraden Anzahl Antwortkategorien ausgehen.

Diese Meinung wird erst 1987 durch das ´Satisficing-Prinzip´ von KROSNICK und ALWIN in Frage gestellt. Die mittlere Antwortkategorie wird demnach auch von den Versuchspersonen gewählt, die auf Grund mangelnder Motivation oder kognitiver Fähigkeiten ihre tatsächliche Meinung nicht durch eine der anderen Antwortkategorien ausdrücken wollen oder können. Die mittlere Kategorie dient in solchen Fällen als opportune Antwortmöglichkeit, die leicht gerechtfertigt werden kann (vgl. KROSNICK, FABRIGAR, 1997, S. 147).

Gefestigt wird das ´Satisficing-Prinzip´ durch Studien, die belegen, dass insbesondere Personen mit geringeren kognitiven Fähigkeiten die mittlere Kategorie wählen (vgl. NARAYAN, KROSNICK, 1996). Ein weiteres Indiz für das ´Satisficing-Prinzip´ liefert eine Arbeit von KROSNICK und SCHUMAN (1988), die belegt, dass Personen die mittlere Kategorie bevorzugt ankreuzen, wenn ihnen die Frage unwichtig ist.

Neben der generellen Frage nach gerad- oder ungeradzahligen Antwortkategorien ist aber auch die exakte Anzahl festzulegen. In Kapitel 4.1.3 wurde bereits der Mappingprozess kurz angesprochen (siehe S. 104). Dieser Prozess beschreibt, welche Schwierigkeiten für die Probanden auftreten, wenn sie ihre Meinung in eine vorgegebene Antwortkategorie übertragen müssen. Beinhaltet eine

Ratingskala zu wenige Antwortkategorien, können die Probanden ihre Meinung nur unpräzise wiedergeben (vgl. ALWIN, KROSNICK, 1991, S. 149). In Abbildung 5.1 ist dieses Problem für eine dreistufige Ratingskala veranschaulicht.

Obwohl die Personen A und B die Frage auf ihrem individuellen psychologischen Kontinuum sehr unterschiedlich beantworten (in Abbildung 5.1 durch den senkrechten Strich gekennzeichnet), kommen sie bei einer 3-stufigen Ratingskala zu identischen Antworten.

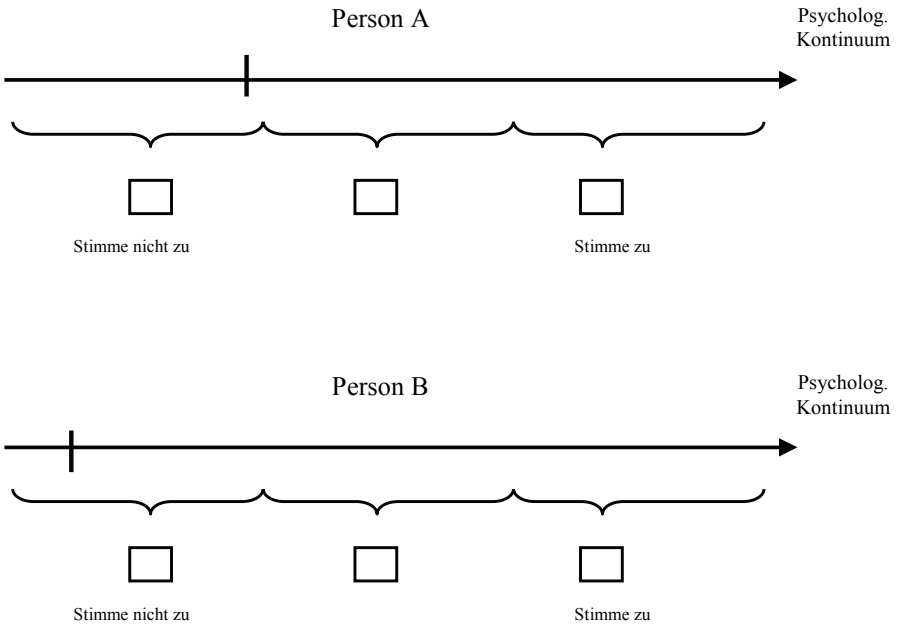


**Abbildung 5.1: Informationsverlust durch die Angabe zu weniger Antwortkategorien**

Die Qualität der Messergebnisse wird aber auch durch zu viele Antwortkategorien negativ beeinflusst, da die Bedeutung der einzelnen Kategorien für die Versuchspersonen unklar wird. Dies hat zur Folge, dass der zur Beantwortung der Frage erforderliche Mappingprozess bei den Probanden sehr inhomogen erfolgt. Deshalb kann die Vorgabe zu vieler Antwortkategorien die Qualität der Ergebnisse ebenfalls negativ beeinflussen (vgl. ALWIN, KROSNICK, 1991, S. 149), wie Abbildung 5.2 verdeutlicht:

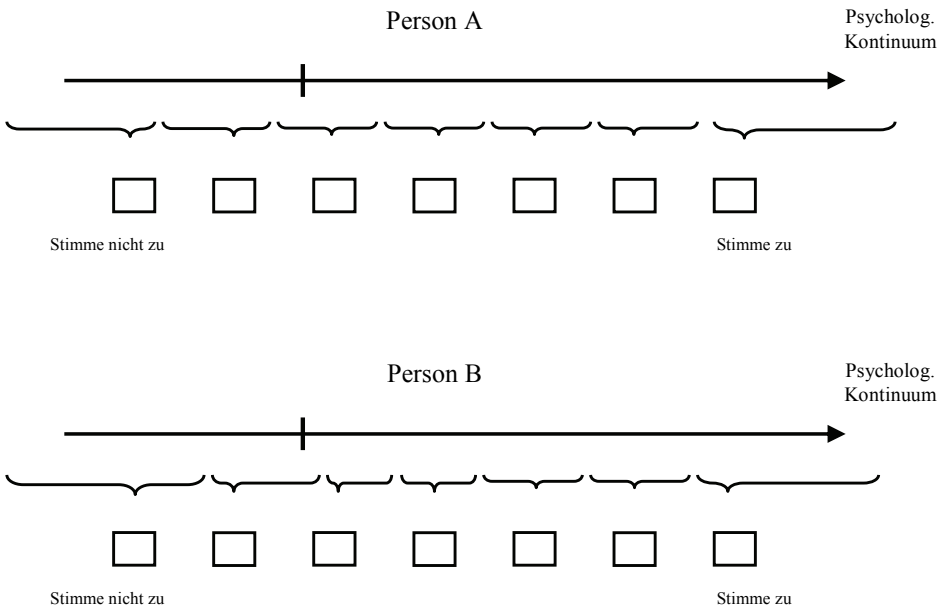


**Abbildung 5.2: Informationsverlust durch die Angabe zu vieler Antwortkategorien**

In Abbildung 5.2 kommen die Personen A und B zur gleichen Einschätzung der Frage bezogen auf ihr psychologisches Kontinuum. Allerdings bilden die beiden Personen diesen Wert auf unterschiedliche Weise in eine der angegebenen Antwortkategorien ab, weshalb sie die Frage nicht identisch beantworten. Dieses Beispiel zeigt, dass auch die Angabe zu vieler Antwortalternativen zu Informationsverlust führen kann. Deshalb hat die richtige Anzahl der Antwortkategorien besonders große Bedeutung für die Qualität der Ratingskala.

Die Ergebnisse der bisher durchgeführten Studien zu diesem Thema sind jedoch sehr uneinheitlich. Deshalb sollen im Folgenden drei aktuelle Studien exemplarisch dargestellt werden. Generell kann festgehalten werden, dass die *Reliabilität* mit zunehmender Zahl der Antwortkategorien steigt. Ab etwa 7 Antwortkategorien führt eine weitere Antwortmöglichkeit jedoch nur noch zu einer geringen Zunahme der Reliabilität. Einige wenige Untersuchungen kommen zum Ergebnis, dass die Reliabilität bei einer geringeren Anzahl als 7 am höchsten ist.

Die *Validität* ist zur Bewertung der Ratingskala das wichtigste Kriterium, da mit ihr angegeben wird, ob tatsächlich gemessen wird, was gemessen werden soll. Wie in Kapitel 3.4.3 (Seite 86 ff.) erläutert, ist die Messung der Validität in der Praxis jedoch mit erheblichen Schwierigkeiten verbunden. Auf Grund dieser

Hindernisse verzichten viele Studien auf die Beurteilung der Messergebnisse mit Hilfe der Validität.

### 5.3.1 Studie von ALWIN, KROSNICK

Bei der Studie von ALWIN und KROSNICK (1991) stehen Einstellungsfragen im Vordergrund. Datengrundlage bilden fünf Paneluntersuchungen mit jeweils etwa 200 Teilnehmern und insgesamt 80 Items mit unterschiedlich vielen Antwortkategorien (vgl. ALWIN, KROSNICK, 1991, S. 155). Paneluntersuchungen sind dadurch gekennzeichnet, dass den Probanden mit zeitlichem Abstand dieselben Fragen erneut zur Beantwortung gegeben werden. Bei den fünf betrachteten Paneluntersuchungen variiert der zeitliche Abstand zweier aufeinander folgender Erhebungen zwischen einem Monat und zwei Jahren (vgl. ALWIN, KROSNICK, 1991, S. 157).

Für jedes Item liegen somit die Daten zweier aufeinander folgender Panelerhebungen vor. Berechnet man für ein Item die Korrelation dieser beiden Befragungen, liefert diese ein Maß für die Reliabilität der Messung. Die Vorgehensweise entspricht der Test-Retest-Methode (S. 81 f.), wie Abbildung 5.3 für eine Paneluntersuchung exemplarisch veranschaulicht:

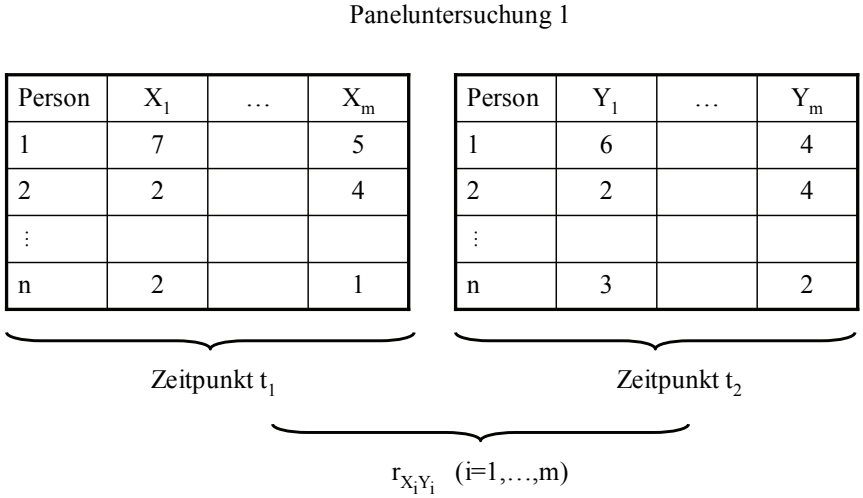


**Abbildung 5.3: Bestimmung der Reliabilität bei der Studie von ALWIN, KROSNICK (1991)**

Die Test-Retest-Methode setzt voraus, dass die Einstellung der Person unverändert bleibt. Deshalb zeigen die Autoren mit Hilfe von Strukturgleichungsmodellen (siehe Kapitel 3.3.3.5) zusätzlich, dass bei den betrachteten Paneluntersuchungen die Einstellung der Personen nahezu konstant bleibt und das Problem veränderter Einstellungen somit vernachlässigbar ist (vgl. ALWIN, KROSNICK, 1991, S. 158f.).

Im nächsten Schritt berücksichtigen die Autoren die unterschiedliche Anzahl der Antwortkategorien der Items. Diese weisen 2, 3, 4, 5, 7 oder 9 Antwortkategorien auf (vgl. ALWIN, KROSNICK, 1991, S. 157). Für jede Anzahl k von Antwortkategorien kann der Mittelwert aller Korrelationskoeffizienten gebildet werden, wie (5.1) zeigt:

$$\bar{r}_{x_i^k y_i^k} = \frac{1}{m_k} \sum_{i=1}^{m_k} r_{x_i^k y_i^k} \tag{5.1}$$

$m_k$: *Anzahl der Items mit k Ausprägungen*

Diese mittlere Korrelation entspricht einem Maß für die durchschnittliche Reliabilität. Diese wiederum dient als Kriterium dafür, welche Zahl der Antwortkategorien zu präferieren ist. In Tabelle 5.1 sind die Mittelwerte der Korrelationen für die unterschiedlichen Skalenbreiten angegeben:

| Anzahl Antwortkategorien | Items | Mittelwert |
|---|---|---|
| 2 | 7 | 0,541 |
| 3 | 22 | 0,477 |
| 4 | 4 | 0,508 |
| 5 | 8 | 0,492 |
| 7 | 10 | 0,572 |
| 9 | 29 | 0,610 |

**Tabelle 5.1: Mittelwerte der Korrelationen (vgl. ALWIN, KROSNICK, 1991, S. 165)**

Bei lediglich zwei Antwortkategorien entsteht mit 0,541 ein relativ hoher Wert für die Reliabilität. Ansonsten liefert die Analyse die von den Autoren erwarteten Ergebnisse, nämlich einen Anstieg der Reliabilität bei Verwendung von mehr Antwortkategorien (vgl. ALWIN, KROSNICK, 1991, S. 164). Die Autoren führen die hohe Reliabilität der Skala mit zwei Antwortkategorien darauf zurück, dass diese lediglich die Richtung und nicht die Stärke einer Einstellung messen. Da

die Richtung einer Einstellung jedoch leichter gemessen werden kann, als die Stärke, wiesen Ratingskalen mit 2 Kategorien eine höhere Reliabilität auf (vgl. ALWIN, KROSNICK, 1991, S. 164). Sie folgern aus diesen Ergebnissen, dass die Qualität einer Messung durch zusätzliche Antwortkategorien steigt.

**Kritik an der Untersuchung von ALWIN, KROSNICK**

Die Autoren vernachlässigen bei ihrer Studie die Validität völlig. Sie merken zwar an, dass auf keinen Fall aus der Reliabilität der Daten auf deren Validität geschlossen werden kann (vgl. ALWIN, KROSNICK, 1991, S. 143), da die Reliabilität lediglich eine notwendige Bedingung für die Validität ist (siehe S. 75). Dennoch unternehmen die Wissenschaftler keinen Versuch, zusätzlich die Validität der Daten zu messen. Es kann also ausschließlich davon ausgegangen werden, dass möglicherweise die Reliabilität leicht zunimmt, wenn mehr Antwortkategorien vorgegeben sind.

Die Studie ist aber noch in einem weiteren Punkt kritisch zu hinterfragen. Die Forscher verwenden Paneluntersuchungen der Vergangenheit, die nicht dafür konzipiert sind, die Reliabilität unterschiedlicher Skalen zu messen. Stattdessen soll mit diesen Skalen möglichst gut die wahre Einstellung der Personen gemessen werden. Der Schluss liegt also nahe, dass die verschiedenen Skalenbreiten nicht zufällig für bestimmte Fragen verwendet werden, sondern systematisch zugewiesen sind. Es kann vermutet werden, dass zielgerichtet Fragestellungen mit bestimmten Skalen abgefragt werden. Die Zunahme der Reliabilität bei einer Erhöhung der Antwortkategorien kann unter Umständen auf diesen Effekt zurückzuführen sein, da Skalen mit vielen Ausprägungen bewusst bei eher einfachen Fragestellungen verwendet werden. Andererseits kann auch die hohe Reliabilität für Skalen mit zwei Antwortkategorien hierauf zurückzuführen sein, weil diese nur bei Fragestellungen eingesetzt werden, die lediglich mit Zu- oder Ablehnung beantwortet werden können.

Die Ergebnisse dieser Studie führen zwar zu interessanten Anhaltspunkten, allerdings erscheint es auf Grund der angesprochenen Kritikpunkte angebracht, weitere Forschungsergebnisse zu betrachten.

## 5.3.2 Studie von TANG, SHAW

Eine im Vergleich zu ALWIN und KROSNICK völlig unterschiedliche Vorgehensweise wählen TANG und SHAW (1999). Bei der Untersuchung müssen die Auskunftspersonen im Rahmen einer Vorstudie angeben, wie relevant insgesamt 30 Thesen (Items) zu einem Thema ihrer Einschätzung nach sind. Die Personen sind in zehn Gruppen mit 2 bis 11 Antwortkategorien eingeteilt (vgl. TANG, SHAW, 1999, S. 257). Der wahre Hintergrund dieses Experiments ist für die Probanden unbekannt (vgl. TANG, SHAW, 1999, S. 258).

Nach jedem der 30 Items werden die Personen gebeten, auf einer Skala von 0-100 zu bewerten, wie sehr sie ihrer eigenen Einschätzung vertrauen (vgl. TANG, SHAW, 1999, S. 257). Der Mittelwert dieser Selbsteinschätzung dient den Autoren als Maßstab, wie gut die jeweilige Ratingskala für den entsprechenden Probanden geeignet ist.

$$\bar{X}_i = \frac{1}{m} \sum_{j=1}^{m} X_{ij} \tag{5.2}$$

*m:* *Anzahl der Items (30)*

$X_{ij}$: *Selbsteinschätzung bezüglich Item j von Person i*

Der Mittelwert über alle Probanden, mit derselben Anzahl Antwortkategorien liefert dann ein Maß für die Güte der Skalenbreite.

$$\bar{X}^k = \frac{1}{n_k} \sum_{i=1}^{n_k} \bar{X}_i^k \tag{5.3}$$

$n_k$: *Anzahl der Personen, die einen Fragebogen mit k Antwortkategorien beantwortet haben*

Die Qualität der Messung wird also mit Hilfe einer Selbsteinschätzung der Versuchspersonen beurteilt (vgl. Kapitel 3.4.4.2). Dieser Ablauf ist in Abbildung 5.4 für die Beurteilung der Ratingskala mit 5 Antwortkategorien schematisch dargestellt (siehe hierzu auch Tabelle 5.2):

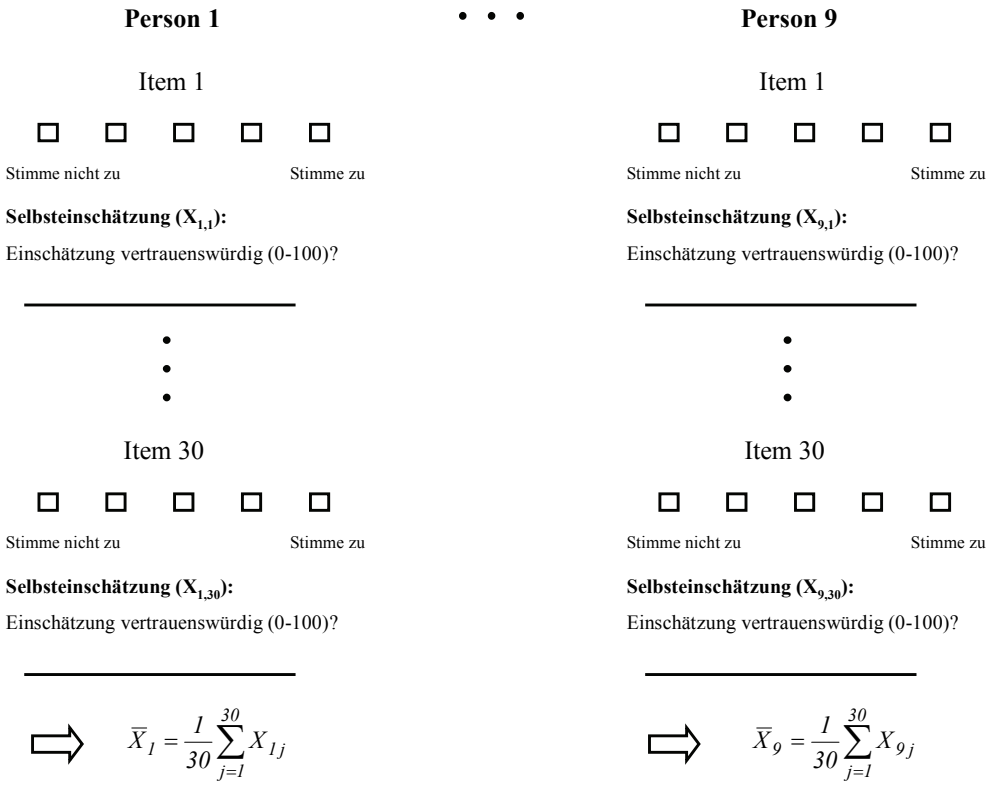


**Abbildung 5.4: Vorgehensweise bei der Untersuchung von TANG, SHAW (1999)**

In Tabelle 5.2 ist für die Vorstudie der Mittelwert der Selbsteinschätzung angegeben und es wird aufgezeigt wie viele Personen jeweils befragt werden:

| Anzahl Antwortkategorien (k) | Personen | $\overline{X}^k$ |
|---|---|---|
| 2 | 12 | 69,7 |
| 3 | 10 | 71,1 |
| 4 | 10 | 72,2 |
| 5 | 9 | 75,4 |
| 6 | 9 | 74,2 |
| 7 | 10 | 72,5 |
| 8 | 8 | 72,8 |
| 9 | 6 | 73,5 |
| 10 | 5 | 68,2 |
| 11 | 7 | 72,7 |

**Tabelle 5.2: Mittelwerte der Selbsteinschätzung bei der Vorstudie (vgl. TANG, SHAW, 1999, S. 260)**

Die Ergebnisse entsprechen in etwa den Erwartungen der Autoren, da das Maximum des Mittelwertes bei 5 bis 6 Kategorien liegt (vgl. TANG, SHAW, 1999, S. 259). Es fällt jedoch auf, dass die Mittelwerte sich kaum unterscheiden und

deshalb eine Varianzanalyse keine signifikanten Unterschiede der Mittelwerte ergibt (vgl. TANG, SHAW, 1999, S. 259).

In der Vorstudie stellen die Autoren fest, dass bei den Ratingskalen mit vielen Antwortkategorien unerwünschte Effekte eintreten. Ein Teil der Probanden verwendet bei allen 30 Items nur einen Ausschnitt der zur Verfügung gestellten Ratingskala, beispielsweise die mittlere Kategorie und die Randkategorien (vgl. TANG, SHAW, 1999, S. 258). Die Autoren argumentieren, dass jene Personen aus der Untersuchung ausgeschlossen werden müssen, die bei allen 30 Items entweder eine der beiden Randkategorien nicht genutzt haben oder zwei aufeinander folgende Kategorien nicht angekreuzt haben (vgl. TANG, SHAW, 1999, S. 258). Bei dieser Vorgehensweise werden hauptsächlich Personen aus der Untersuchung ausgeschlossen, die sehr viele Antwortkategorien zur Verfügung haben. Dies erklärt auch, weshalb in der Vorstudie die Anzahl der befragten Personen mit steigender Zahl der Antwortkategorien abnimmt (vgl. Tabelle 5.2).

Neben dem Ausschluss von Personen gelangen die Autoren zusätzlich zur Ansicht, dass bei einer Person nicht alle 30 Items zur Berechnung des Mittelwertes der Selbsteinschätzung herangezogen werden sollen. Viel mehr soll die Selbsteinschätzung zu einem Item nur dann berücksichtigt werden, wenn eine der beiden Randkategorien angekreuzt ist (vgl. TANG, SHAW, 1999, S. 260).

Da diese Vorgehensweise bei der später vorgenommenen Kritik des Ansatzes im Mittelpunkt steht, soll sie nochmals kurz wiederholt werden:

1. Ausschluss von Personen, die bei allen 30 Items entweder
   - keine Randkategorie ankreuzen oder
   - zwei aufeinander folgende Kategorien nicht benutzen.
2. Ausschluss von Items bei der Berechnung des Mittelwertes der Selbsteinschätzung, wenn bei der zu Grunde liegenden Ratingskala keine Randkategorie angekreuzt ist.

In der auf dieser Weise durchgeführten Hauptstudie erhält man die in Tabelle 5.3 dargestellten Mittelwerte der Selbsteinschätzung. Eine Varianzanalyse kommt zum Ergebnis, dass die Mittelwertunterschiede signifikant sind.

| Anzahl Antwortkategorien (k) | Personen | $\overline{X}^k$ |
|---|---|---|
| 2 | 10 | 0,663 |
| 3 | 10 | 0,713 |
| 4 | 11 | 0,740 |
| 5 | 13 | 0,783 |
| 6 | 10 | 0,839 |
| 7 | 10 | 0,788 |
| 8 | 10 | 0,754 |
| 9 | 11 | 0,800 |
| 10 | 11 | 0,742 |
| 11 | 9 | 0,797 |

**Tabelle 5.3: Mittelwerte der Selbsteinschätzung bei der Hauptstudie (vgl. TANG, SHAW, 1999, S. 261)**

### Kritik an der Untersuchung von TANG, SHAW

Ein erster Kritikpunkt an der Studie von TANG und SHAW ist die Entfernung bestimmter Fragebögen aus der Untersuchung. Insbesondere bei Ratingskalen mit vielen Antwortkategorien werden Personen ausgeschlossen, während bei 2- oder 3-stufigen Ratingskalen faktisch alle Personen analysiert werden. Besonders kritisch an dieser Ungleichbehandlung der einzelnen Ratingskalen ist die Tatsache, dass der Ausschluss von Personen nicht zufällig erfolgt. Gerade solche Personen, die auf Grund ihres Ankreuzverhaltens zeigen, dass ihnen Ratingskalen mit vielen Antwortkategorien keinen zusätzlichen Nutzen stiften, werden bei der weiteren Untersuchung vernachlässigt. Es ist daher zu erwarten, dass die Ergebnisse der Selbsteinschätzung zu Gunsten der Ratingskalen mit vielen Antwortkategorien verzerrt sind. Dieser Punkt ist deshalb von Bedeutung, weil die Autoren angeben, dass insgesamt etwa 30% der befragten Personen nicht berücksichtigt werden. Es muss deshalb davon ausgegangen werden, dass die Beeinflussung der Ergebnisse sehr groß ist.

Ein weiterer Kritikpunkt ist darin zu sehen, dass lediglich solche Items zur Berechnung der Mittelwerte herangezogen werden, bei denen die Probanden eine Extremkategorie angekreuzt haben. Dies führt ebenfalls zu einer Verzerrung der Ergebnisse zu Gunsten der Ratingskalen mit vielen Antwortkategorien. Denn bei lediglich 2 Antwortkategorien entspricht jede Antwortkategorie einer Randkategorie. Dementsprechend werden alle Items analysiert. Bei 11 Antwortkategorien wird ein Item dagegen nur dann berücksichtigt, wenn die

Person von ihrer Ansicht vollkommen überzeugt ist und daher eine Randkategorie wählt. Es ist davon auszugehen, dass sie in diesem Fall auch bei der Selbsteinschätzung angibt, von ihrer Antwort überzeugt zu sein.

In Anbetracht dieser Schwachpunkte der Studie von TANG, SHAW (1999) müssen die Ergebnisse kritisch hinterfragt werden. Welche Anzahl Antwortkategorien bei einer Ratingskala geeignet ist, kann mit Hilfe dieser Studie nicht beantwortet werden.

### 5.3.3 Studie von PRESTON, COLMAN

In der Studie von PRESTON und COLMAN werden den Probanden Fragen gestellt, die sich in folgende vier Gruppen einteilen lassen (vgl. PRESTON, COLMAN, 2000, S. 4):

1. Jede Person muss im Fragebogen zunächst auf einer Skala von 1-100 angeben, wie zufrieden sie mit der Servicequalität beim letzten Besuch eines Restaurants oder Warenhauses war (vgl. PRESTON, COLMAN, 2000, S. 4).

2. Anschließend werden fünf Teilaspekte der Servicequalität mit Hilfe von Ratingskalen (2 bis 11 Antwortkategorien) oder einer Skala mit 1-100 Punkten bewertet (vgl. PRESTON, COLMAN, 2000, S. 4). Die Personen werden somit in 11 unterschiedliche Gruppen eingeteilt.

3. Außerdem müssen die Probanden die verschiedenen Skalenbreiten dahingehend bewerten (1-100 Punkte), wie leicht sie anwendbar sind, wie schnell sie anwendbar sind und wie sehr sie den Personen die Möglichkeit einräumen, ihre Einstellung adäquat auszudrücken (vgl. PRESTON, COLMAN, 2000, S. 5).

4. Ein bis drei Wochen später müssen die Probanden den Fragebogen erneut ausfüllen.

Die Beurteilung der Qualität erfolgt auf Basis verschiedener Kriterien:

- Die *Validität* der Ergebnisse liefert ein Vergleich der fünf Teilaspekte (2) mit der Gesamtbewertung (1) der Servicequalität.

- Ein Vergleich der Ergebnisse der ersten Befragung (2) mit der Befragung ein bis drei Wochen später (4) und die interne Konsistenz der 5 Teilaspekte der Servicequalität (2) liefert die *Reliabilität* der Messung.

- Auf Basis der Gesamtbeurteilung (1) können die Auskunftspersonen in zwei Gruppen eingeteilt werden, eine Gruppe mit positiver Gesamteinschätzung und eine Gruppe mit negativer Gesamteinschätzung. Mit Hilfe dieser Gruppeneinteilung kann die *Diskriminierungsfähigkeit* der einzelnen Ratingskalen (2) überprüft werden.

- Die *Selbsteinschätzung* der Qualität der verschiedenen Ratingskalen (3) dient abschließend als weiterer Maßstab.

**Überprüfung der Validität**

Bei der Beurteilung der *Validität* dient die Gesamtbewertung mit der Skala von 1-100 als externes Kriterium. Somit messen die Autoren die Übereinstimmungsvalidität (siehe Seite 87) zwischen einem durch Aggregation der fünf Einzelurteile erzeugten Wert (X) und der Gesamtbewertung (Y). In Tabelle 5.4 ist die Validität, also die Korrelation zwischen aggregierten Einzelbewertungen (2) und Gesamtbewertung (1), der verschiedenen Ratingskalen angegeben.

| Anzahl Antwortkategorien | $r_{XY}$ |
|---|---|
| 2 | 0,83 |
| 3 | 0,82 |
| 4 | 0,85 |
| 5 | 0,87 |
| 6 | 0,88 |
| 7 | 0,87 |
| 8 | 0,87 |
| 9 | 0,89 |
| 10 | 0,87 |
| 11 | 0,88 |
| 1-100 | 0,89 |

**Tabelle 5.4: Korrelation der Ratingskalen mit dem Gesamturteil (vgl. PRESTON, COLMAN, 2000, S. 7)**

Durch eine Zunahme der Antwortkategorien erhöht sich zunächst die Validität. Sie bleibt auch bei sehr vielen Antwortkategorien konstant hoch und nimmt nicht ab, so bald ein Schwellenwert überschritten ist.

**Überprüfung der Reliabilität**

Die *Reliabilität* wird sowohl mit Hilfe von Cronbach´s α (siehe Seite 84) als auch mit der Test-Retest-Methode (siehe Seite 82) gemessen. Bei der Bestimmung von Cronbach´s α wird die interne Konsistenz der fünf Teilaspekte der Servicequalität (2) in Abhängigkeit der verwendeten Ratingskala herangezogen. Die Test-Retest-Methode vergleicht die Antworten zu den fünf Teilaspekten im ersten Fragebogen (2) mit der zweiten Befragung (4) hinsichtlich ihrer Konsistenz (vgl. PRESTON, COLMAN, 2000, S. 5). So kann für jedes der 5 Items die Korrelation zwischen erster und zweiter Befragung bestimmt werden. Der Mittelwert dieser 5 Korrelationskoeffizienten dient als Maß für die Reliabilität der betreffenden Ratingskala. Die Ergebnisse für Cronbach´s α und die Test-Retest-Methode sind in Tabelle 5.5 dargestellt.

| Anzahl Antwortkategorien | Cronbach´s α | Test-Retest |
|---|---|---|
| 2 | 0,81 | 0,88 |
| 3 | 0,79 | 0,86 |
| 4 | 0,82 | 0,89 |
| 5 | 0,82 | 0,91 |
| 6 | 0,83 | 0,92 |
| 7 | 0,85 | 0,93 |
| 8 | 0,85 | 0,94 |
| 9 | 0,85 | 0,94 |
| 10 | 0,85 | 0,93 |
| 11 | 0,86 | 0,92 |
| 1-100 | 0,85 | 0,90 |

**Tabelle 5.5: Reliabilität gemessen mit der Test-Retest-Methode und Cronbach´s α (vgl. PRESTON, COLMAN, 2000, S. 6)**

Die Reliabilität nimmt auch bei dieser Studie mit einer Erhöhung der Antwortkategorien zunächst zu. Sobald eine gewisse Kategorienzahl erreicht ist, nimmt die Reliabilität dann wieder ab. Bei Cronbach´s α ist die Abnahme der Reliabilität ab einem bestimmten Schwellenwert jedoch nicht zu erkennen. Hier verlaufen die Werte ab etwa 7 Kategorien weitestgehend konstant. Ansonsten ist

kein wesentlicher Unterschied zwischen der Test-Retest-Methode und Cronbach´s α ersichtlich.

**Überprüfung der Diskriminierungsfähigkeit**

Neben der Reliabilität und der Validität betrachten die Autoren zusätzlich die *Diskriminierungsfähigkeit* (siehe Seite 91 f.) der Skalen. Dazu teilen sie die Versuchspersonen in zwei Gruppen ein. Gruppe A hat bei der Gesamtbewertung der Servicequalität (1) einen Wert zwischen 0 und 25 angegeben, Gruppe B einen Wert über 80 (vgl. PRESTON, COLMAN, 2000, S. 7). Damit erhalten die Autoren eine Gruppe mit sehr unzufriedenen Personen und eine Gruppe mit sehr zufriedenen Personen.

Anschließend werden die fünf Teilaspekte der Servicequalität (2) dahingehend überprüft, ob sie zwischen diesen beiden Gruppen diskriminieren. Um die Mittelwertunterschiede verschiedener Skalenbreiten vergleichbar zu machen, formen die Autoren die angegebenen Skalenwerte folgendermaßen um (vgl. PRESTON, COLMAN, 2000, S. 7):

$$x_{Norm} = \frac{(x-1)}{(k-1)} \cdot 100 \qquad (5.4)$$

*x:* *Von der Person angegebener Skalenwert*

*k:* *Anzahl der Antwortkategorien (Skalenbreite)*

Diese Umformung bewirkt, dass die Skalenwerte bei allen Skalenbreiten Werte zwischen 0 und 100 annehmen, wie Tabelle 5.6 veranschaulicht:

| Anzahl Antwortkategorien | X | $X_{Norm}$ |
|---|---|---|
| 2 | 1;2 | 0;100 |
| 3 | 1;2;3 | 0;50;100 |
| 4 | 1;2;3;4 | 0;33;66;100 |
| ⋮ | ⋮ | ⋮ |
| 1-100 | 1;2;…;100 | 0;1,01;…;100 |

**Tabelle 5.6: Normierung der Skalenwerte unterschiedlicher Antwortkategorien**

Nach dieser Umformung ist ein Vergleich der Mittelwertunterschiede zwischen Personengruppe A und B für verschiedene Skalenbreiten möglich. Dennoch verwenden die Autoren nicht den Mittelwertunterschied als Maßstab für die

Diskriminierungsfähigkeit, sondern berechnen den Wert der Teststatistik V für einen Zweistichproben-t-Test auf einen signifikanten Mittelwertunterschied zwischen den beiden Personengruppen (vgl. BAMBERG, BAUR, 2002, S. 193):

$$V = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2} \cdot \frac{n_1 + n_2}{n_1 \cdot n_2}}} \tag{5.5}$$

Die Autoren erläutern nicht, weshalb sie die Teststatistik an Stelle des Mittelwertunterschiedes verwenden. Auf diesen Punkt wird im Rahmen der Kritik der Studie noch ausführlicher eingegangen. In Tabelle 5.7 sind die Ergebnisse für die verschiedenen Skalenbreiten festgehalten:

| Anzahl Antwortkategorien | Teststatistik |
|---|---|
| 2 | 19,2 |
| 3 | 16,8 |
| 4 | 18,6 |
| 5 | 20,8 |
| 6 | 21,5 |
| 7 | 20,8 |
| 8 | 22,2 |
| 9 | 23,7 |
| 10 | 22,3 |
| 11 | 23,0 |
| 0-100 | 23,4 |

**Tabelle 5.7: Diskriminierungsfähigkeit der Ratingskalen (vgl. PRESTON, COLMAN, 2000, S. 7)**

Erneut steigt die Qualität der Ergebnisse durch eine Zunahme der Antwortkategorien deutlich an. Wie zuvor bei der Validität bereits beobachtet, ist auch kein Punkt erkennbar, ab dem die Werte wieder sinken würden.

### Überprüfung der Skalenbreiten mit Hilfe der Selbsteinschätzung

Abschließend werten die Autoren die direkten Fragen nach der *Selbsteinschätzung* (siehe Seite 93 f.) aus. Tabelle 5.8 fasst diese Ergebnisse zusammen:

| Anzahl Antwortkategorien | Einfachheit | Schnelligkeit | Exaktheit |
|---|---|---|---|
| 2 | 78,6 | 86,6 | 17,8 |
| 3 | 81,4 | 86,8 | 40,0 |
| 4 | 82,0 | 85,5 | 52,0 |
| 5 | 83,7 | 85,1 | 63,7 |
| 6 | 81,3 | 84,5 | 63,4 |
| 7 | 82,3 | 83,5 | 69,0 |
| 8 | 81,5 | 83,1 | 68,8 |
| 9 | 81,0 | 82,1 | 72,9 |
| 10 | 83,2 | 82,9 | 76,0 |
| 11 | 76,7 | 77,8 | 73,1 |
| 0-100 | 74,1 | 70,6 | 79,3 |

**Tabelle 5.8: Bewertung der Ratingskalen durch die Probanden (vgl. PRESTON, COLMAN, 2000, S. 10)**

Bei der Einfachheit der Fragestellung ist keine eindeutige Tendenz zu erkennen. Anscheinend haben Personen insbesondere mit sehr wenigen (2) und sehr vielen Antwortkategorien (11) sowie der Einschätzung auf einer Skala von 0-100 erhebliche Probleme. Dagegen lässt sich für die Schnelligkeit der Handhabung die Aussage ableiten, dass diese mit steigender Kategorienzahl abnimmt. Durch eine Erhöhung der Kategorienzahl erhöht sich jedoch die Möglichkeit für die Probanden, ihre tatsächliche Einstellung auszudrücken (vgl. PRESTON, COLMAN, 2000, S. 10). Unter diesem Gesichtspunkt ist die Nutzung möglichst vieler Antwortkategorien empfehlenswert.

In der Zusammenfassung ihrer Ergebnisse kommen die Autoren zu dem Schluss, dass es für die Verwendung der weit verbreiteten Ratingskala mit 5 Antwortkategorien keine Rechtfertigung gibt (vgl. PRESTON, COLMAN, 2000, S. 13). Stattdessen empfehlen die Verfasser die Benutzung von Ratingskalen mit 7, 9 oder 10 Kategorien.

### Kritische Würdigung der Untersuchung von PRESTON, COLMAN

Die Autoren verwenden unterschiedliche Kriterien zur Messung der *Reliabilität*, dennoch liefern beide Kennzahlen vergleichbare Ergebnisse. Darüber hinaus erscheint die Vorgehensweise bei der Berechnung der Kennzahlen ebenfalls unproblematisch. Die Interpretation der Ergebnisse für die Reliabilität ist daher uneingeschränkt möglich.

Die Selbsteinschätzung der Probanden ist kein objektives Maß zur Beurteilung der Qualität. Sie kann Aussagen zur Validität jedoch unter Umständen untermauern. Bei der Überprüfung der *Validität* verwenden die Autoren ein externes Kriterium (1), das im selben Fragebogen wie die zu überprüfenden Einzelbewertungen (2) gemessen wird. Diese Vorgehensweise wird allgemein als unproblematisch angesehen (vgl. PRESTON, COLMAN, 2000, S. 6). Wesentlich kritischer ist jedoch die Tatsache, dass bei der Bestimmung der Übereinstimmungsvalidität die Messung des externen Kriteriums (1) ebenfalls mit Hilfe einer Skala (0-100) erfolgt. Es muss davon ausgegangen werden, dass die Messung des externen Kriteriums selbst fehlerhaft ist. Wie in Kapitel 4.1.3 bereits erläutert und in Abbildung 5.2 zu Anfang dieses Kapitels explizit dargestellt, übertragen die Versuchspersonen ihre persönlichen Einstellungen auf völlig unterschiedliche Weise auf unterschiedliche Antwortkategorien. Dies gilt wahrscheinlich auch für die Bewertung mit Hilfe eines Score-Wertes.

Die Ergebnisse zur *Diskriminierungsfähigkeit* müssen ebenfalls kritisch hinterfragt werden. Die Autoren liefern keine Erklärung, weshalb sie den Wert der Teststatistik an Stelle des Mittelwertunterschiedes verwenden. Bei der Teststatistik muss zusätzlich der Einfluss der Standardabweichung der beiden Gruppen ($s_1$ und $s_2$) berücksichtigt werden. Durch die Transformation der Skalenwerte weist die 2-stufige Ratingskala die Ausprägungen 0 und 100 auf. Eine 7-stufige Ratingskala besitzt dagegen die Ausprägungen 0, 17, 33, 50, 67, 83 und 100. Es ist daher davon auszugehen, dass die Standardabweichung bei der 2-stufige Ratingskala höher ist, als bei der 7-stufigen Ratingskala. Dies hat zur Folge, dass Ratingskalen mit vielen Ausprägungen bei der Teststatistik tendenziell besser abschneiden, als jene mit nur sehr wenigen Antwortkategorien.

Als Schwachstellen der Arbeit erweisen sich somit die Aussagen über die Validität und Diskriminierungsfähigkeit der Kennzahlen. Daher erscheinen in diesem Bereich weitere Untersuchungen sinnvoll.

# Kapitel 6

# Empirische Studie zur Qualität von Ratingskalen

Eine Übersicht über die bestehende Literatur kann nicht eindeutig klären, wie eine Ratingskala konkret auszugestalten ist. Insbesondere die Frage nach der idealen Anzahl der Antwortkategorien kann durch die bisher veröffentlichten Arbeiten nicht abschließend beantwortet werden. Dieses Problem wird deshalb im Mittelpunkt des sechsten Kapitels stehen.

- Kapitel 6.1 beschreibt hierfür zunächst den Aufbau der eigenen empirischen Studie.
- Kapitel 6.2 stellt dann methodisches Vorgehen und die Ergebnisse der Untersuchung vor.

## 6.1 Aufbau der empirischen Studie

Um den Aufbau der eigens angelegten Studie nachvollziehen zu können, wird zunächst nochmals ausführlich die Problemstellung der Untersuchung erläutert, die deren Notwendigkeit beschreibt (Kapitel 6.1.1). Mit Hilfe der Problemstellung können dann die Anforderungen an die empirische Studie formuliert werden. Diese Anforderungen münden schließlich in einer konkreten Zielsetzung für die Untersuchung (Kapitel 6.1.2). Gemäß dieser Zielsetzung wird anschließend ein Fragebogen ausgearbeitet. Das Untersuchungsdesign ist dabei

so gewählt, dass alle in Kapitel 6.1.2 aufgeworfenen Fragen beantwortet werden können (Kapitel 6.1.3).

### 6.1.1 Problemstellung der eigenen Studie

Wie die Kapitel 3.1 bis 3.3 gezeigt haben, sind Ratingskalen in den Wirtschafts- und Sozialwissenschaften von großer Wichtigkeit. Diese enorme Bedeutung bei der Messung im wirtschaftswissenschaftlichen Umfeld leitet sich daraus ab, dass die zu messende Größe nicht direkt erfasst werden kann. Es existieren deshalb viele verschiedene Skalierungsverfahren, die dieses Problem zu lösen versuchen und dabei auf die Ratingskala zurückgreifen.

Zur Beurteilung der Qualität einer Ratingskala sind die in Kapitel 3.4 vorgestellten Methoden heranzuziehen. Besondere Bedeutung haben dabei Reliabilität und Validität der Messung. Wie die Ausführungen in Kapitel 4 gezeigt haben, können psychologische Prozesse für Ungenauigkeiten bei der Messung verantwortlich sein. Beide Aspekte – wie die Qualität gemessen wird und welche Fehlerquellen bei der Messung mit einer Ratingskala überhaupt bestehen – müssen berücksichtigt werden, wenn Empfehlungen für die Ausarbeitung einer Ratingskala gegeben werden.

Eine der zentralen Entscheidungen bei der Erstellung einer Ratingskala ist, wie viele Antwortkategorien diese umfassen soll. Die in Kapitel 5 vorgestellten Arbeiten zur optimalen Anzahl der Antwortkategorien stimmen darin überein, dass die *Reliabilität* einer Ratingskala steigt, wenn zusätzliche Antwortkategorien berücksichtigt werden. Dieser Effekt nimmt deutlich ab, wenn bereits mehr als 7 Antwortkategorien vorhanden sind. Ein Teil der Autoren kommt sogar zu dem Ergebnis, dass die Reliabilität ab einer gewissen Zahl der Antwortkategorien durch eine zusätzliche Kategorie sinkt.

Die übereinstimmenden Ergebnisse für die Reliabilität sollen in dieser Studie nicht in Frage gestellt werden. Allerdings misst die Reliabilität nur das Fehlen eines zufälligen Fehlers ($X_R$). Wie Abbildung 3.22 bereits gezeigt hat, kann eine perfekt reliable Messung dennoch systematische Fehler ($X_S$) aufweisen. Deshalb soll Bild 2 dieser Abbildung hier erneut aufgegriffen werden:

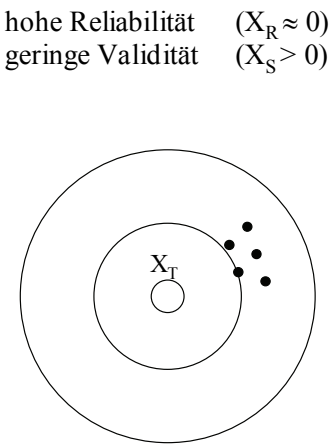


**Abbildung 6.1: Geringe Validität einer Messung trotz hoher Reliabilität**

Gerade die Ausführungen in Kapitel 4 zu inhomogenen Mappingprozessen der Versuchspersonen legen aber den Schluss nahe, dass es zu systematischen Fehlern bei der Messung mit Hilfe einer Ratingskala kommen kann. In Abbildung 5.2 wurde bereits verdeutlicht, wie durch die Angabe zu vieler Antwortalternativen systematische Fehler entstehen können. Daher soll auch diese Abbildung nochmals aufgeführt werden:

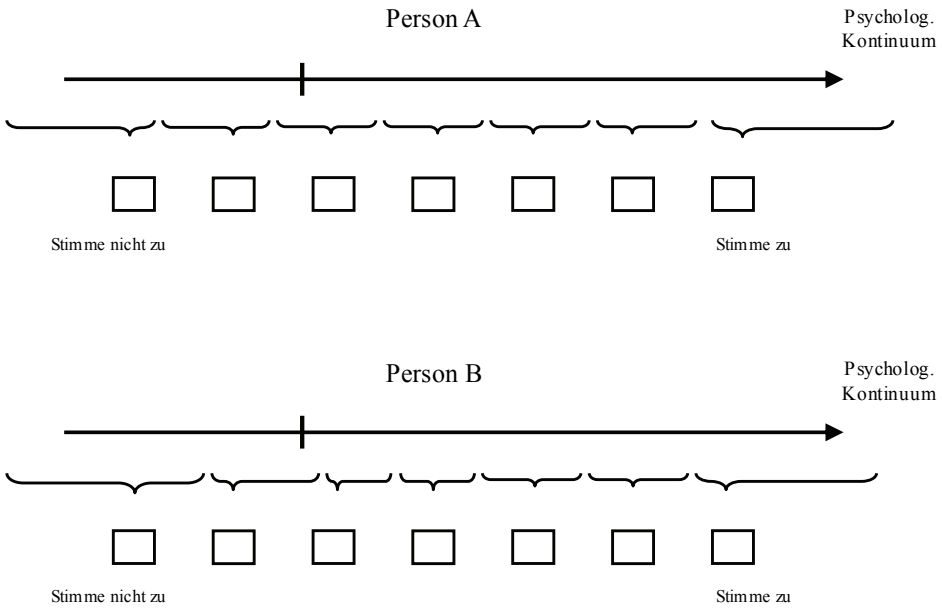


**Abbildung 6.2: Entstehung systematischer Fehler durch die Angabe zu vieler Antwortkategorien**

Die *Validität* als Kennzahl für die Existenz eines systematischen Fehlers wird bei vielen Untersuchungen nicht herangezogen. Dies liegt wahrscheinlich an der Schwierigkeit, die Validität adäquat zu messen. Dazu ist entweder ein umfangreiches theoretisches Modell zur Messung der Konstruktvalidität (vgl. Kapitel 3.4.3.2) oder die Messung eines externen Kriteriums zur Bestimmung der Kriteriumsvalidität (vgl. Kapitel 3.4.3.1) erforderlich.

Die Wissenschaftler PRESTON und COLMAN (2000) ermöglichen durch die Messung eines externen Kriteriums die Beurteilung der Validität. Jedoch messen die Autoren dieses externe Kriterium mit einer der Ratingskala vergleichbaren Skala von 1-100 Punkten. Daher müssen die Versuchspersonen auch bei der Beantwortung der Frage nach dem externen Kriterium den Abbildungsprozess durchlaufen. Es ist deshalb davon auszugehen, dass die Messung des externen Kriteriums nicht fehlerfrei erfolgt. Insofern besitzen die Ergebnisse zur Validität der Ratingskalen nur eingeschränkte Aussagekraft.

Die bisherigen Forschungsergebnisse müssen jedoch noch in einem weiteren Punkt kritisch hinterfragt werden. Keine der aktuellen Studien berücksichtigt nämlich die Inhomogenität der Abbildungsprozesse als mögliche Ursache für ungenaue Messungen mit einer Ratingskala. Die Kenntnis dieser Fehlerquelle könnte es ermöglichen, den Mappingprozess der Probanden zu homogenisieren und die Güte der Messung somit zu steigern.

Die Kritik an den bestehenden Ansätzen kann in zwei Punkten zusammengefasst werden:

- Die Messung der Validität erfolgt mit ungeeigneten Methoden oder wird völlig ignoriert.
- Die Kenntnisse der in Kapitel 4 dargestellten psychologischen Prozesse werden nicht dafür genutzt, die Messung mit einer Ratingskala zu verbessern.

Aus den Kritikpunkten lassen sich die beiden zu erörternden Fragestellungen für die eigene empirische Studie direkt ableiten:

1. Wie kann die Validität einer Ratingskala geeignet gemessen werden?

2. Kann die Messung mit Hilfe einer Ratingskala verbessert werden, wenn die Inhomogenität der Mappingprozesse bekannt ist?

Ein besonderes Augenmerk wird dabei auf die Frage gelegt, wie viele Antwortkategorien bei einer Ratingskala optimal sind.

### 6.1.2 Zielsetzung der empirischen Studie

Gemäß der Fragestellungen der Problemstellung lassen sich daraus für die eigene Studie folgende Zielsetzungen ableiten:

1. Die empirische Studie soll erforschen, welche Anzahl der Antwortkategorien unter dem Gesichtspunkt der Validität optimal ist.

2. Die empirische Studie soll belegen, dass inhomogene Mappingprozesse existieren.

3. Die empirische Studie soll Möglichkeiten aufzeigen, wie durch die Identifikation unterschiedlicher Mappingstrategien die Qualität einer Ratingskala gesteigert werden kann.

### 6.1.3 Untersuchungsdesign

Bei der empirischen Untersuchung werden 573 Studenten der wirtschaftswissenschaftlichen Fakultät der Universität Augsburg befragt. Die Durchführung der Erhebung erfolgt innerhalb einer Woche in verschiedenen Lehrveranstaltungen.

Im Rahmen dieser Arbeit soll die Messung der Kriteriumsvalidität erfolgen. Diese hat gegenüber der Konstruktvalidität den Vorteil, dass kein theoretisches Modell aufgestellt werden muss. Die Überprüfung der Kriteriumsvalidität setzt allerdings voraus, dass ein *externes Kriterium* möglichst fehlerfrei gemessen wird. Mit diesem externen Kriterium können verschiedene *Ratingskalen* beurteilt werden. Abschließend werden die Probanden zusätzlich nach einer Einschätzung eines *durchschnittlichen Studenten* befragt. Der Fragebogen gliedert sich somit in drei Teilbereiche:

- Messung des externen Kriteriums

- Messung mit Hilfe der Ratingskala
- Messung der Einschätzung eines durchschnittlichen Studenten

Im ersten Schritt werden die Probanden direkt nach der (metrischen) Ausprägung verschiedener Merkmale befragt, die als externes Kriterium zur Beurteilung der Ratingskalen dienen. Wichtig ist dabei, dass bei der Beantwortung der Frage kein Abbildungsprozess durchlaufen werden muss. In Tabelle 6.1 sind diese Fragen dargestellt:

| | |
|---|---|
| *Wie viele Stunden bereiten Sie, abgesehen von der direkten Prüfungsvorbereitung, die Vorlesungen wöchentlich vor bzw. nach?* | |
| *Wie viele Stunden lernen Sie in der Vorbereitungszeit auf die Prüfungen durchschnittlich am Tag?* | |
| *Wie viele Stunden jobben Sie regelmäßig während des Semesters pro Woche?* | |
| *Wie hoch ist ungefähr Ihr monatliches Einkommen (in Euro)* ***vor*** *Abzug fixer Kosten (Lohn, Taschengeld, BAföG, ...)?* | |
| *Wie viel Geld (in Euro) steht Ihnen monatlich zur freien Verfügung?* | |
| *An wie vielen Tagen (pro Semester) engagieren Sie sich außerhalb des Studiums im engeren Sinn für die Universität (StuRa, AStA, ...)?* | |
| *Wie viele Stunden arbeiten Sie pro Woche ehrenamtlich (Rotes Kreuz, Sportverein, ...)?* | |
| *Wie weit entfernt von der Universität wohnen Sie (km)?* | |
| *Wie viel Zeit (in Stunden) nehmen Sie sich wöchentlich für Hobbys (Sport, Musik, ...)?* | |
| *Wie viel Miete bezahlen Sie monatlich (Euro)?* | |
| *Wie viele Quadratmeter Wohnfläche stehen ausschließlich Ihnen zur Verfügung?* | |
| *Wie viele Geschwister haben Sie?* | |
| *Wie viele Jahre beträgt der Altersunterschied Ihrer Eltern?* | |
| *Wie viele Stunden Schlaf benötigen Sie, um sich fit zu fühlen?* | |
| *Wie groß sind Sie (in cm)?* | |

**Tabelle 6.1: Befragung nach dem externen Kriterium**

Diese Fragen müssen von allen Versuchspersonen beantwortet werden. Die Antworten bilden das externe Kriterium, mit dessen Hilfe die Validität der ebenfalls erhobenen Ratingskalen bestimmt werden kann. Wichtig ist dabei, dass

es bei der Erhebung dieses externen Kriteriums zu keinen Verzerrungen kommt. Bei der Beantwortung der Frage könnte der in Kapitel 4.5.2.3 beschriebene Rundungseffekt auftreten (siehe Seite 132 f.). Bei den zu erwartenden Antworten ist jedoch davon auszugehen, dass dieser Effekt bei den meisten Fragen vernachlässigt werden kann.

In einem zweiten Schritt müssen die Versuchspersonen zu denselben Themenstellungen wie bei der Messung des externen Kriteriums auch mit Hilfe einer Ratingskala Stellung beziehen. Dazu werden die Probanden nach ihrer Einschätzung gefragt, wie sich ihre Merkmalsausprägung des externen Kriteriums im Vergleich zu einer durchschnittlichen Person verhält. Dieser Teil des Fragebogens ist nicht bei allen Versuchspersonen identisch. Es werden 3 Gruppen gebildet, die sich hinsichtlich der Anzahl der Antwortkategorien unterscheiden:

- Gruppe 1: 3 Antwortkategorien
- Gruppe 2: 5 Antwortkategorien
- Gruppe 3: 7 Antwortkategorien

Die Fragen sind in Tabelle 6.2 exemplarisch für einen Fragebogen mit 5 Antwortkategorien dargestellt. Eingeleitet wurde dieser Teil des Fragebogens mit folgender Formulierung:

*Wenn Sie sich mit **anderen Studenten vergleichen**, wie sehr würden Sie dann folgenden Aussagen zustimmen? Beziehen Sie sich dabei bitte auf Ihre **derzeitige Situation**!*

| | *Stimme voll und ganz zu* | *Stimme eher zu* | *Teils teils* | *Stimme eher nicht zu* | *Stimme über-haupt nicht zu* |
|---|---|---|---|---|---|
| *Ich verwende während des Semesters viel Zeit auf die Vor- und Nachbereitung von Vorlesungen* | ☐ | ☐ | ☐ | ☐ | ☐ |
| *Mein tägliches Lernpensum ist in der Vorbereitungszeit auf die Prüfungen größer als bei anderen Studenten* | ☐ | ☐ | ☐ | ☐ | ☐ |
| *Während des Semesters jobbe ich überdurchschnittlich viel* | ☐ | ☐ | ☐ | ☐ | ☐ |
| *Für einen Studenten ist mein monatliches Einkommen (Lohn, Taschengeld, ...) relativ hoch (vor Abzug der Fixkosten)* | ☐ | ☐ | ☐ | ☐ | ☐ |
| *Für einen Studenten habe ich sehr viel Geld zur freien Verfügung* | ☐ | ☐ | ☐ | ☐ | ☐ |
| *Mein ehrenamtliches Engagement für die Universität (StuRa, AStA, ...) ist groß* | ☐ | ☐ | ☐ | ☐ | ☐ |
| *Mein ehrenamtliches Engagement außerhalb der Universität (Rotes Kreuz, Sportverein, ...) ist groß* | ☐ | ☐ | ☐ | ☐ | ☐ |
| *Ich habe eine weite Anfahrt von meinem Wohnort zur Universität* | ☐ | ☐ | ☐ | ☐ | ☐ |
| *Ich verwende überdurchschnittlich viel Zeit für Hobbys (Sport, Musik, ...)* | ☐ | ☐ | ☐ | ☐ | ☐ |
| *Ich bezahle eine vergleichsweise hohe Miete* | ☐ | ☐ | ☐ | ☐ | ☐ |
| *Ich habe sehr viel Wohnraum für mich allein* | ☐ | ☐ | ☐ | ☐ | ☐ |
| *Im Vergleich zu anderen habe ich viele Geschwister* | ☐ | ☐ | ☐ | ☐ | ☐ |
| *Der Altersunterschied meiner Eltern ist groß* | ☐ | ☐ | ☐ | ☐ | ☐ |
| *Ich benötige viel Schlaf, um mich fit zu fühlen* | ☐ | ☐ | ☐ | ☐ | ☐ |
| *Ich bin überdurchschnittlich groß* | ☐ | ☐ | ☐ | ☐ | ☐ |

**Tabelle 6.2: Messung mit Hilfe von Ratingskalen**

Die Antworten der Probanden in diesem Teil der Studie werden in eine äquidistante Skala mit 3, 5 bzw. 7 verschiedenen Skalenwerten überführt. In Abbildung 6.3 ist die Zuordnung der Skalenwerte für die 3-stufige Ratingskala dargestellt. Bei der 5- und 7-stufigen Skala erfolgt die Zuweisung analog. Bei der 5-stufigen Ratingskala werden den Versuchspersonen also Werte zwischen 1 und 5 zugeordnet.

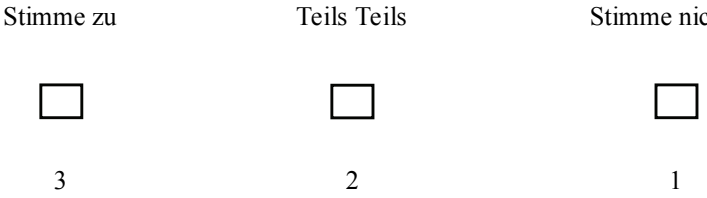


**Abbildung 6.3: Zuweisung von Skalenwerten bei der 3-stufigen Ratingskala**

Bei der Beantwortung mit Hilfe der Ratingskalen ist die Problemstellung für die Probanden zweigeteilt:

- Die Versuchspersonen benötigen eine Vorstellung über die Höhe der Merkmalsausprägung einer durchschnittlichen Person. Sie müssen also zum Beispiel abschätzen, wie viel Miete ein durchschnittlicher Student bezahlt. Diesen Wert der durchschnittlichen Person müssen sie dann mit der eigenen Merkmalsausprägung vergleichen und entscheiden, wie sehr sie der Aussage zustimmen, überdurchschnittlich viel Miete zu bezahlen. Hier sind das *Abrufen und die Beurteilung relevanter Information* durch den Probanden erforderlich.

- Die Probanden müssen diese Einschätzung auf einer Ratingskala zum Ausdruck bringen. Dazu ist es erforderlich, die individuelle Ausprägung der metrischen Variable, im Vergleich zu einer durchschnittlichen Person, in einer Antwortkategorie geeignet *abzubilden*.

Aus dieser Vielschichtigkeit der für den Probanden auftretenden Probleme resultiert, dass Verzerrungen der Messung mit Hilfe der Ratingskala nicht zwangsläufig auf den Abbildungsprozess zurückzuführen sind. Stattdessen kann auch eine inhomogene Einschätzung eines durchschnittlichen Studenten die Ursache sein. Im Hinblick auf Ziel 2 der empirischen Studie – der Nachweis, dass inhomogene Mappingstrategien erhebliche Auswirkungen auf die Qualität der Messergebnisse einer Ratingskala haben können – müssen die Auswirkung inhomogener *Informationsbeurteilung* und inhomogener *Abbildungsprozesse* getrennt voneinander betrachtet werden können.

Deshalb müssen die Versuchspersonen bei solchen Merkmalen, die eine besonders starke Schwankung bei der Beurteilung einer Durchschnittsperson nahe

legen, zusätzlich beurteilen, welche Ausprägung sie bei einer durchschnittlichen Person vermuten. Diese Fragen sind in Tabelle 6.3 dargestellt.

Eingeleitet werden die Fragen mit folgender Formulierung:

*Beantworten Sie jetzt bitte die Fragen nochmals: Geben Sie jetzt **Ihre Einschätzung** ab, welche Werte für einen **durchschnittlichen Studenten** gelten:*

| | |
|---|---|
| *Wie hoch ist ungefähr das monatliche Einkommen in Euro **vor** Abzug fixer Kosten (Lohn, Taschengeld, BAföG, ...)?* | |
| *Wie viel Geld (in Euro) steht monatlich zur freien Verfügung?* | |
| *Wie viel Miete wird monatlich bezahlt (Euro)?* | |
| *Wie weit ist die Entfernung von der Universität zum Wohnort (km)?* | |
| *Wie viele Quadratmeter Wohnfläche stehen ausschließlich zur eigenen Verfügung?* | |
| *Wie viele Stunden wird während des Semesters pro Woche ungefähr gejobbt?* | |
| *Wie viel Zeit (in Stunden) wird wöchentlich für Hobbys (Sport, Musik, ...) verwendet?* | |
| *Wie viele Stunden pro Woche wird ehrenamtlich gearbeitet (Rotes Kreuz, Sportverein, ...)?* | |
| *An wie vielen Tagen (pro Semester) erfolgt ein Engagement für die Universität außerhalb des Studiums im engeren Sinn (StuRa, AsTa, etc.)?* | |
| *Wie viele Stunden bereitet man, abgesehen von der direkten Prüfungsvorbereitung, die Vorlesungen wöchentlich vor bzw. nach?* | |
| *Wie viele Stunden lernt man in der Vorbereitungszeit auf die Prüfungen durchschnittlich am Tag?* | |

**Tabelle 6.3: Fragen nach der Einschätzung eines durchschnittlichen Studenten**

Diesen Teil des Fragebogens müssen alle Probanden der Studie unabhängig von der Anzahl der Antwortkategorien beantworten.

Neben den bisher aufgeführten Fragen müssen die Probanden auch ihr Geschlecht mitteilen. Dies ermöglicht separate Auswertungen für Frauen und Männer. Bei der Frage nach der Körpergröße ist dies erforderlich, weil die Beurteilung der eigenen Körpergröße vom Geschlecht abhängig ist. So wird z.B. eine 180 cm große Frau der Aussage überdurchschnittlich groß zu sein eher zustimmen, als ein gleich großer Mann. Der komplette Fragebogen ist im Anhang angegeben.

Die erhobenen Daten bilden die entscheidende Grundlage für das Verständnis der Ergebnisse der empirischen Studie, daher sollen diese zum Abschluss dieses Kapitels nochmals anschaulich in Tabelle 6.4 dargestellt werden. In Analogie zu Kapitel 3 soll das externe Kriterium mit $y_j$ und die auf die Validität zu überprüfende Ratingskala mit $x_j$ bezeichnet werden. Die Frage nach der Einschätzung der Durchschnittsperson wird mit $z_j$ benannt.

| Themenstellung | Metrische Ausprägung (Externes Kriterium) | Messung mit Ratingskala | Einschätzung der Durchschnittsperson |
|---|---|---|---|
| Vorlesung | $y_1$ | $x_1$ | $z_1$ |
| Lernstunden | $y_2$ | $x_2$ | $z_2$ |
| Jobben | $y_3$ | $x_3$ | $z_3$ |
| Einkommen | $y_4$ | $x_4$ | $z_4$ |
| Geld | $y_5$ | $x_5$ | $z_5$ |
| Universität | $y_6$ | $x_6$ | $z_6$ |
| Ehrenamt | $y_7$ | $x_7$ | $z_7$ |
| Entfernung | $y_8$ | $x_8$ | $z_8$ |
| Hobbys | $y_9$ | $x_9$ | $z_9$ |
| Miete | $y_{10}$ | $x_{10}$ | $z_{10}$ |
| Wohnfläche | $y_{11}$ | $x_{11}$ | $z_{11}$ |
| Geschwister | $y_{12}$ | $x_{12}$ | --- |
| Altersunterschied | $y_{13}$ | $x_{13}$ | --- |
| Schlaf | $y_{14}$ | $x_{14}$ | --- |
| Größe-Männer | $y_{15}$ | $x_{15}$ | --- |
| Größe-Frauen | $y_{16}$ | $x_{16}$ | --- |

**Tabelle 6.4: Bezeichnung der erhobenen Daten**

## 6.2 Methodisches Vorgehen und Ergebnisse der eigenen Studie

Nach der Erhebung der Daten müssen diese mit unterschiedlichen Methoden ausgewertet werden. Zusammen mit dem Untersuchungsdesign unterstreichen die verwendeten Methoden, welche zusätzlichen Erkenntnisse die eigene Studie im Vergleich zu den in Kapitel 5 vorgestellten Arbeiten liefert. Des Weiteren unterstützen die Ergebnisse der Untersuchung die ausformulierten Ziele:

- Es ist zunächst eine Bewertung der verschiedenen Ratingskalen erforderlich (Kapitel 6.2.1).

- Anschließend wird in Kapitel 6.2.2 die Existenz inhomogener Abbildungsprozesse belegt.
- Kapitel 6.2.3 verdeutlicht schließlich, inwieweit eine Identifikation der Abbildungsprozesse die Qualität der Messung erhöhen kann.

## 6.2.1 Bewertung der verschiedenen Ratingskalen

Gemäß **Zielsetzung 1** der empirischen Studie erfolgt ein Vergleich der verschiedenen Ratingskalen.

- In Kapitel 6.2.1.1 wird die Kriteriumsvalidität der verschiedenen Ratingskalen überprüft.
- Die Beurteilung der Diskriminierungsfähigkeit der Ratingskalen wird in Kapitel 6.2.1.2 vorgenommen.
- Kapitel 6.2.1.3 fasst die Ergebnisse schließlich zusammen und kommt so zu einer Aussage, welche der betrachteten Ratingskalen zu präferieren ist.

### 6.2.1.1 Validität der Ratingskalen

Als Maßstab zur Bewertung der Validität dient die Korrelation zwischen interessierender Ratingskala ($x_j$) und externem Kriterium ($y_j$).

$$V_{x_j} = r_{x_j y_j} \tag{6.1}$$

$V_{x_j}$ : *Validität des Messinstruments* $x_j$

$r_{x_j y_j}$ : *Korrelation zwischen Messinstrument* $x_j$ *und externem Kriterium* $y_j$

In Anlehnung an die Ausführungen in Kapitel 3.4.3.1 stellt Abbildung 6.4 die Vorgehensweise zur Bestimmung der Validität schematisch dar:

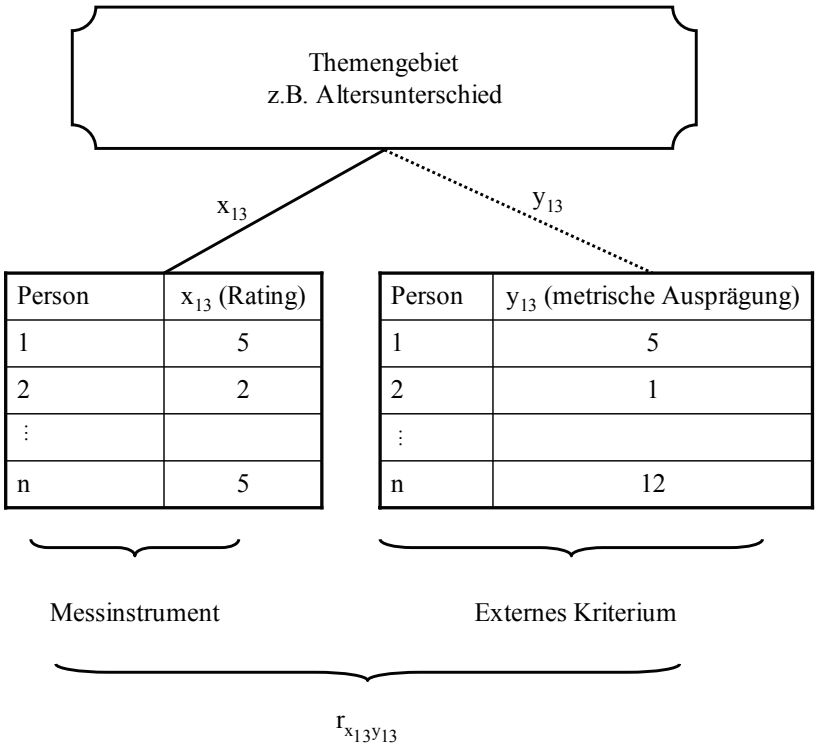


**Abbildung 6.4: Beurteilung der Ratingskalen mit Hilfe der Kriteriumsvalidität**

Die Berechung der Korrelation zwischen zwei Größen $x_j$ und $y_j$ soll mit den in Kapitel 3.4.1.1 vorgestellten Korrelationsmaßen erfolgen (siehe Seite 77-80):

- Bravais-Pearson
- **Spearman**
- **Kendalls τ-b**
- Kendalls τ-c

Es ermöglichen jedoch nicht alle Korrelationsmaße gleichermaßen eine geeignete Beurteilung der verschiedenen Ratingskalen. Dafür sind zwei Problemfelder verantwortlich:

- Ausreißer bei der Messung des externen Kriteriums
- Bindungen bei der Messung mit Hilfe der Ratingskalen

Die *Ausreißerproblematik* führt dazu, dass der Bravais-Pearson-Korrelations-koeffizient zur Beurteilung der Validität der Ratingskalen nicht herangezogen

wird. Bei seiner Berechnung fließen nämlich die tatsächlichen Merkmalsausprägungen des externen Kriteriums ein. Bei der (metrischen) Messung des externen Kriteriums sind allerdings vereinzelte Ausreißer zu erwarten, zum Beispiel Personen, die über ein sehr hohes monatliches Einkommen verfügen. Diese besonders hohen oder niedrigen Werte können nicht in eine Ratingskala übertragen werden, weshalb der Bravais-Pearson-Korrelationskoeffizient stark abnimmt. Diese Eigenschaft ist deshalb so kritisch, weil Ausreißer nicht bei allen Teilstichproben einheitlich auftreten, da bei 3-, 5- und 7-stufiger Ratingskala unterschiedliche Personen befragt werden. Somit wäre die Beurteilung der Ratingskalen auf Basis dieses Koeffizienten davon abhängig, bei welcher Teilstichprobe zufällig weniger Ausreißer beim externen Kriterium auftreten.

Die ordinalen Korrelationsmaße (Spearman und Kendalls τ-b bzw. τ-c) benötigen dagegen nicht die tatsächliche Merkmalsausprägung, sondern lediglich den Rang bzw. die Anzahl konkordanter und diskordanter Objektpaare. Deshalb sind sie in dieser Studie besser geeignet, die Qualität der Ratingskalen zu beurteilen.

Das Auftreten von *Bindungen*, also identischen Merkmalsausprägungen bei verschiedenen Versuchspersonen, ist ein zweiter Problembereich der empirischen Studie. Bindungen müssen bei der Messung mit Hilfe der Ratingskalen auftreten, da lediglich 3, 5 oder 7 unterschiedliche Merkmalsausprägungen möglich sind. Im Sinne einer möglichst exakten Messung sollten Bindungen bei den Ratingskalen ($x_j$) nur dann bestehen, wenn die beiden Probanden auch beim externen Kriterium ($y_j$) identische Merkmalsausprägungen vorweisen. Andernfalls führt die Messung mit Hilfe der Ratingskala zu Informationsverlust.

Dementsprechend müssen die Korrelationsmaße folgende Eigenschaft vorweisen: Eine Bindung bei lediglich einem Merkmal ($x_j$ oder $y_j$) muss dazu führen, dass die Korrelation sinkt.

Der Korrekturfaktor von Kendalls τ-c sorgt allerdings dafür, dass Bindungen bei nur einem der beiden Merkmale nicht zu einer Verringerung der Korrelation führen. Da umso mehr Bindungen zu erwarten sind, je weniger Antwortalternativen bestehen, würde Kendalls τ-c Ratingskalen mit wenigen Antwortkategorien bevorzugen. Deshalb soll auch auf eine Bewertung der Ratingskalen mit Hilfe dieser Kennzahl verzichtet werden.

Bei der Interpretation der absoluten Korrelationswerte für Kendalls τ-b und den Rangkorrelationskoeffizienten nach Spearman ist ihre Eigenschaft zu berücksichtigen, dass Bindungen bei nur einem Merkmal eine Verringerung der Korrelationskoeffizienten nach sich ziehen. Denn bei der Messung mit Hilfe der Ratingskala treten zwangsläufig Bindungen auf, die bei der Messung des externen Kriteriums nicht vorhanden sind. Daher sind die absoluten Werte für die Korrelation vergleichsweise niedrig.

In Tabelle 6.5 sind die Korrelationen zwischen den drei Ratingskalen und dem jeweiligen externen Kriterium gemessen mit *Kendalls τ-b* dargestellt. Die relativ niedrigen absoluten Korrelationswerte können teilweise auf die hohe Anzahl von Bindungen bei den Ratingskalen ($x_j$) zurückgeführt werden. Ein Vergleich der drei betrachteten Skalen liefert allerdings interessante Ergebnisse.

Die Ratingskala mit 5 Kategorien stellt unter dem Gesichtspunkt der Validität häufig die beste Ratingskala dar. Im Vergleich zur Ratingskala mit 3 Kategorien liefert sie bei 12 von 16 Vergleichen eine höhere Korrelation mit dem externen Kriterium. Stellt man sie der 7-stufigen Skala gegenüber, weist sie bei 11 von 16 Vergleichen einen größeren Zusammenhang auf. Interessant ist die Betrachtung von 3- und 7-stufiger Ratingskala. Bei 10 von 16 Merkmalen ist die Korrelation der 3-stufigen Skala höher als die Korrelation der 7-stufigen Skala. Dieses Resultat widerspricht den Ergebnissen von PRESTON und COLMAN (2000) und zahlreicher weiterer Studien, die 7-stufige Ratingskalen empfehlen.

| | 3-stufige Skala | 5-stufige Skala | 7-stufige Skala |
|---|---|---|---|
| Vorlesung ( $r_{x_1y_1}^{\tau-b}$ ) | 0,387 | 0,561 | 0,420 |
| Lernstunden ( $r_{x_2y_2}^{\tau-b}$ ) | 0,279 | 0,297 | 0,160 |
| Jobben ( $r_{x_3y_3}^{\tau-b}$ ) | 0,566 | 0,632 | 0,554 |
| Einkommen ( $r_{x_4y_4}^{\tau-b}$ ) | 0,388 | 0,392 | 0,255 |
| Geld ( $r_{x_5y_5}^{\tau-b}$ ) | 0,355 | 0,418 | 0,300 |
| Universität ( $r_{x_6y_6}^{\tau-b}$ ) | 0,453 | 0,342 | 0,365 |
| Ehrenamt ( $r_{x_7y_7}^{\tau-b}$ ) | 0,635 | 0,540 | 0,521 |
| Entfernung ( $r_{x_8y_8}^{\tau-b}$ ) | 0,382 | 0,483 | 0,328 |
| Hobbys ( $r_{x_9y_9}^{\tau-b}$ ) | 0,447 | 0,361 | 0,385 |
| Miete ( $r_{x_{10}y_{10}}^{\tau-b}$ ) | 0,523 | 0,665 | 0,528 |
| Wohnfläche ( $r_{x_{11}y_{11}}^{\tau-b}$ ) | 0,500 | 0,406 | 0,448 |
| Geschwister ( $r_{x_{12}y_{12}}^{\tau-b}$ ) | 0,673 | 0,707 | 0,760 |
| Altersunterschied ( $r_{x_{13}y_{13}}^{\tau-b}$ ) | 0,484 | 0,502 | 0,584 |
| Schlaf ( $r_{x_{14}y_{14}}^{\tau-b}$ ) | 0,500 | 0,510 | 0,403 |
| Größe-Männer ( $r_{x_{15}y_{15}}^{\tau-b}$ ) | 0,590 | 0,616 | 0,595 |
| Größe-Frauen ( $r_{x_{16}y_{16}}^{\tau-b}$ ) | 0,458 | 0,567 | 0,542 |

**Tabelle 6.5: Korrelation (Kendalls τ-b) der Ratingskala mit dem externen Kriterium**

Diese Erkenntnisse sind nicht vom gewählten Korrelationskoeffizienten abhängig, denn auch bei der Verwendung des Rangkorrelationskoeffizienten nach *Spearman* ändert sich an der Rangfolge der Skalen nur in wenigen Fällen etwas zu Gunsten der 7-stufigen Ratingskala. Die absoluten Werte der Korrelationskoeffizienten sind jedoch für alle Skalen höher, wie Tabelle 6.6 verdeutlicht:

| | 3-stufige Skala | 5-stufige Skala | 7-stufige Skala |
|---|---|---|---|
| Vorlesung ( $r^{SP}_{x_1y_1}$ ) | 0,465 | 0,672 | 0,533 |
| Lernstunden ( $r^{SP}_{x_2y_2}$ ) | 0,339 | 0,377 | 0,210 |
| Jobben ( $r^{SP}_{x_3y_3}$ ) | 0,625 | 0,733 | 0,655 |
| Einkommen ( $r^{SP}_{x_4y_4}$ ) | 0,478 | 0,489 | 0,317 |
| Geld ( $r^{SP}_{x_5y_5}$ ) | 0,438 | 0,516 | 0,403 |
| Universität ( $r^{SP}_{x_6y_6}$ ) | 0,479 | 0,374 | 0,412 |
| Ehrenamt ( $r^{SP}_{x_7y_7}$ ) | 0,683 | 0,625 | 0,628 |
| Entfernung ( $r^{SP}_{x_8y_8}$ ) | 0,454 | 0,601 | 0,435 |
| Hobbys ( $r^{SP}_{x_9y_9}$ ) | 0,543 | 0,451 | 0,495 |
| Miete ( $r^{SP}_{x_{10}y_{10}}$ ) | 0,619 | 0,791 | 0,658 |
| Wohnfläche ( $r^{SP}_{x_{11}y_{11}}$ ) | 0,615 | 0,525 | 0,590 |
| Geschwister ( $r^{SP}_{x_{12}y_{12}}$ ) | 0,722 | 0,765 | 0,824 |
| Altersunterschied ( $r^{SP}_{x_{13}y_{13}}$ ) | 0,555 | 0,592 | 0,680 |
| Schlaf ( $r^{SP}_{x_{14}y_{14}}$ ) | 0,574 | 0,598 | 0,485 |
| Größe-Männer ( $r^{SP}_{x_{15}y_{15}}$ ) | 0,710 | 0,736 | 0,742 |
| Größe-Frauen ( $r^{SP}_{x_{16}y_{16}}$ ) | 0,540 | 0,676 | 0,674 |

**Tabelle 6.6: Korrelation (Rangkorrelationskoeffizient) der Ratingskala mit dem externen Kriterium**

Bei der Frage nach der *Größe* der Versuchspersonen sind deutliche Unterschiede bei den Ergebnissen für Frauen und Männer festzustellen (vgl. Tabelle 6.5, 6.6). Eine Ursache für die schlechteren Ergebnisse bei den Frauen könnte sein, dass Männer die Frage ´Ich bin überdurchschnittlich groß´ einheitlich so interpretieren, dass sie ein überdurchschnittlich großer Mann sind. Analog versteht auch ein Teil der Frauen die Frage so, dass sie eine überdurchschnittlich große Frau sind. Einige Frauen beziehen diese Frage aber vielleicht auf beide Geschlechter, weshalb sie die Aussage überdurchschnittlich groß zu sein verneinen, obwohl sie im Vergleich zu anderen Frauen eher groß sind. Damit wären die deutlich niedrige-

ren Korrelationen beim Merkmal *Größe-Frauen* im Vergleich zu *Größe-Männer* erklärbar.

Für die Korrelationsmaße aus den Tabellen 6.5 und 6.6 kann der Mittelwert für jede Ratingskala gebildet werden. In Tabelle 6.7 sind diese Mittelwerte angegeben:

| | 3-stufige Skala | 5-stufige Skala | 7-stufige Skala |
|---|---|---|---|
| $\bar{r}_{x_j y_j}^{\tau-b}$ | 0,476 | 0,500 | 0,447 |
| $\bar{r}_{x_j y_j}^{SP}$ | 0,552 | 0,595 | 0,546 |

**Tabelle 6.7: Mittelwert der Korrelationskoeffizienten über alle Merkmale**

Die Mittelwerte der Korrelationskoeffizienten unterstreichen die bereits getroffenen Aussagen: Die 5-stufige Skala ist am besten geeignet, die Information des metrischen Merkmals abzubilden. Am schlechtesten schneidet die 7-stufige Skala ab, während die 3-stufige Skala zwar schlechter als die 5-stufige Skala ist, jedoch besser als die 7-stufige Skala.

Neben der deskriptiven Auswertung der Korrelationskoeffizienten sind induktive Aussagen wünschenswert. Bei der induktiven Analyse von Korrelationskoeffizienten existieren vor allem Tests auf Unkorreliertheit zweier Merkmale. Für normalverteilte Zufallsvariablen sind auch Tests über die Gleichheit der Korrelationskoeffizienten von je zwei Merkmalspaaren möglich (vgl. HARTUNG, ELPELT, 1992, S. 159). Die ratingskalierten Merkmale können jedoch nicht normalverteilt sein. Daher ist die Konstruierung eines geeigneten statistischen Tests nahezu unmöglich.

Um dennoch Aussagen über die Signifikanz treffen zu können, sollen die Korrelationskoeffizienten selbst als Realisierungen einer Zufallsstichprobe aufgefasst werden. Für die 3-stufige Ratingskala können 16 Korrelationen beobachtet werden, ebenso für die 5- und 7-stufige Ratingskala. Signifikanzaussagen werden mit Hilfe des Wilcoxon-Vorzeichenrangtests getroffen[2]. Getestet wird also, ob der Median der Korrelationen der 3-, 5- und 7-stufigen Ratingskalen übereinstimmt oder nicht (vgl. BOSCH, 1998, S.691 ff.).

[2] Der Wilcoxon-Vorzeichenrangtest setzt voraus, dass die einzelnen Stichprobenrealisierungen – also die einzelnen Korrelationskoeffizienten – alle derselben Verteilung entstammen. Der Test wird durchgeführt, da diese Voraussetzung eher erfüllt ist, als eine Normalverteilung der einzelnen Ratingskalen.

$$H_0 : \mu_{x_j y_j}^{med}(3) = \mu_{x_j y_j}^{med}(5)$$

$$H_0 : \mu_{x_j y_j}^{med}(3) = \mu_{x_j y_j}^{med}(7)$$

$$H_0 : \mu_{x_j y_j}^{med}(5) = \mu_{x_j y_j}^{med}(7)$$

$\mu_{x_j y_j}^{med}(3)$: *Median aller Korrelationen der 3-stufigen Ratingskala*

$\mu_{x_j y_j}^{med}(5)$: *Median aller Korrelationen der 5-stufigen Ratingskala*

$\mu_{x_j y_j}^{med}(7)$: *Median aller Korrelationen der 7-stufigen Ratingskala*

Beim Test zweier Ratingskalen wird zunächst die Differenz für jedes Korrelationspaar bestimmt. Anschließend wird über den Betrag dieser Differenz eine Rangfolge gebildet. Tabelle 6.8 veranschaulicht die Vorgehensweise beim Vergleich der 5- und 7-stufigen Ratingskala mit Kendalls τ-b:

| | $r_{x_j y_j}^{\tau-b}(5)$ | $r_{x_j y_j}^{\tau-b}(7)$ | $r_{x_j y_j}^{\tau-b}(5) - r_{x_j y_j}^{\tau-b}(7)$ | $\mathrm{Rg}\left\lvert r_{x_j y_j}^{\tau-b}(5) - r_{x_j y_j}^{\tau-b}(7)\right\rvert$ |
|---|---|---|---|---|
| Vorlesung | 0,561 | 0,420 | 0,141 | 15 |
| Lernstunden | 0,297 | 0,160 | 0,137 | 12 |
| Jobben | 0,632 | 0,554 | 0,078 | 8 |
| Einkommen | 0,392 | 0,255 | 0,137 | 13 |
| Geld | 0,418 | 0,300 | 0,118 | 11 |
| Universität | 0,342 | 0,365 | -0,023 | 3 |
| Ehrenamt | 0,540 | 0,521 | 0,019 | 1 |
| Entfernung | 0,483 | 0,328 | 0,155 | 16 |
| Hobbys | 0,361 | 0,385 | -0,024 | 4 |
| Miete | 0,665 | 0,528 | 0,137 | 13 |
| Wohnfläche | 0,406 | 0,448 | -0,042 | 6 |
| Geschwister | 0,707 | 0,760 | -0,053 | 7 |
| Altersunterschied | 0,502 | 0,584 | -0,082 | 9 |
| Schlaf | 0,510 | 0,403 | 0,107 | 10 |
| Größe-Männer | 0,616 | 0,595 | 0,021 | 2 |
| Größe-Frauen | 0,567 | 0,542 | 0,025 | 5 |

**Tabelle 6.8: Berechnung der Teststatistik für den Wilcoxon-Vorzeichenrangtest**

Die Teststatistik entspricht dann der Rangsumme der Korrelationen mit positiven bzw. negativen Differenzen. So ergibt sich für den Vergleich von 5- und 7-stufiger Ratingskala mit Hilfe von Kendalls τ-b eine Teststatistik von 107 (Summe der Rangplätze aller positiven Differenzen).

In Tabelle 6.9 sind die positiven Rangsummen für die entsprechenden Vergleiche aufgelistet. Zum Vergleich ist in Klammern auch die negative Rangsumme angegeben. Die mit einem ´*´ gekennzeichnete Rangsumme ist bei einem Vorzeichenrangtest zu einer Irrtumswahrscheinlichkeit von 5% signifikant.

| | Rangsumme: $r_{x_j y_j}^{\tau-b}(3) - r_{x_j y_j}^{\tau-b}(7)$ | Rangsumme: $r_{x_j y_j}^{\tau-b}(5) - r_{x_j y_j}^{\tau-b}(3)$ | Rangsumme: $r_{x_j y_j}^{\tau-b}(5) - r_{x_j y_j}^{\tau-b}(7)$ |
|---|---|---|---|
| Kendalls τ-b | 97 (39) | 92 (44) | 107 (29)* |
| Rangkorrelationskoeffizient | 73 (63) | 99 (37) | 100 (36) |

**Tabelle 6.9: Teststatistiken des Vorzeichenrangtests nach Wilcoxon**

Die Teststatistik für den Vorzeichenrangtest unterstreicht die bereits festgestellten Tendenzen. Die 5-stufige Ratingskala ist am besten, während die 7-stufige Ratingskala am schlechtesten abschneidet. Allerdings ist lediglich der Unterschied zwischen 5- und 7-stufiger Ratingskala signifikant.

### 6.2.1.2 Diskriminierungsfähigkeit der Ratingskalen

Neben der Berechnung der Korrelation als Maßstab für die Validität der Ratingskalen wird bei empirischen Studien teilweise auch die Diskriminierungsfähigkeit verschiedener Ratingskalen als Qualitätskriterium herangezogen (vgl. PRESTON, COLMAN, 2000, S. 7). Bei der Diskriminierungsfähigkeit handelt es sich streng genommen nicht um eine Methode zur Messung der Validität, wie die Ausführungen in Kapitel 3.4.4.1 gezeigt haben. Sie soll deshalb in dieser Studie lediglich als zusätzliche Bestätigung für die mit Hilfe der Korrelationsmaße gewonnenen Erkenntnisse dienen.

Zur Berechnung der Diskriminierungsfähigkeit werden die Versuchspersonen anhand der Ausprägungen beim externen Kriterium ($y_j$) in zwei Gruppen eingeteilt. Eine Personengruppe mit sehr hohen (Gruppe 1) und eine Gruppe mit

sehr niedrigen Werten (Gruppe 2). Anschließend werden die Mittelwertunterschiede der beiden Gruppen bezogen auf die verschiedenen Ratingskalen ($x_j$) bestimmt. Ein besonders großer Mittelwertunterschied spricht dann dafür, dass die Ratingskala gut diskriminiert. Zur Sicherstellung der Vergleichbarkeit der Mittelwerte verschiedener Skalenbreiten werden zuvor die Skalenwerte, wie bei der Studie von PRESTON und COLMAN (2000) beschrieben, transformiert (siehe S. 148).

$$x_{ij}^{Norm} = \frac{(x_{ij} - 1)}{(k - 1)} \cdot 100 \qquad (6.2)$$

$x_{ij}$: *Von der Person i beim Merkmal $x_j$ angegebener Skalenwert (i=1,..., n; j= 1,..., 16)*

$k$: *Anzahl der Antwortkategorien (Skalenbreite) bei $x_{ij}$*

In Abbildung 6.5 ist die Vorgehensweise bei der Bestimmung der Diskriminierungsfähigkeit für die Frage nach dem Altersunterschied exemplarisch für acht Probanden dargestellt:

| Person | $x_{13}^{Norm}$ | $y_{13}$ | |
|---|---|---|---|
| 1 | 100 | 13 | Gruppe 1 |
| 2 | 100 | 10 | |
| 3 | 50 | 8 | |
| 4 | 100 | 5 | |
| 5 | 0 | 4 | |
| 6 | 50 | 2 | |
| 7 | 0 | 2 | Gruppe 2 |
| 8 | 0 | 1 | |

| Gruppe | $\bar{x}_{13}^{Norm}$ |
|---|---|
| 1 | 100 |
| 2 | 0 |

**Abbildung 6.5: Beispiel für die Bestimmung der Diskriminierungsfähigkeit**

Die Einteilung der Gruppen auf Basis des externen Kriteriums ($y_j$) erfolgt so, dass sie jeweils etwa 25% der Personen der ursprünglichen Stichprobe umfassen. Von dieser Vorgehensweise wird bei den Themengebieten *Jobben, Uni* und *Ehrenamt* abgewichen, da mehr als die Hälfte der Personen bei der Messung von $y_j$ den Wert 0 angeben. Daher bilden hier alle Personen, die einen Wert ungleich 0 angeben Gruppe 1 und die übrigen Probanden Gruppe 2.

Die Mittelwertunterschiede zwischen den beiden Gruppen können mit Hilfe eines approximativen Zweistichproben-Gaußtests überprüft werden (vgl. BAMBERG, BAUR, 2002, S. 193). Die Nullhypothese lautet in diesem Fall:

$H_0$: $\mu_{Gr1} = \mu_{Gr2}$

Zu einer Irrtumswahrscheinlichkeit von 95% erweisen sich alle in Tabelle 6.10 dargestellten Mittelwertunterschiede als signifikant:

| | Mittelwertdifferenz (Gruppe 1 - Gruppe2) 3-stufige Skala | Mittelwertdifferenz (Gruppe 1 - Gruppe2) 5-stufige Skala | Mittelwertdifferenz (Gruppe 1 - Gruppe2) 7-stufige Skala |
|---|---|---|---|
| Vorlesung | 30,2 | 40,8 | 36,9 |
| Lernstunden | 36,0 | 28,3 | 17,5 |
| Jobben | 43,8 | 44,7 | 36,2 |
| Einkommen | 51,4 | 35,0 | 16,7 |
| Geld | 46,5 | 34,7 | 32,7 |
| Universität | 49,6 | 42,7 | 43,3 |
| Ehrenamt | 59,3 | 43,4 | 42,2 |
| Entfernung | 52,1 | 55,2 | 34,3 |
| Hobbys | 50,4 | 24,8 | 30,9 |
| Miete | 58,9 | 45,5 | 29,4 |
| Wohnfläche | 76,5 | 44,1 | 45,3 |
| Geschwister | 68,5 | 61,9 | 65,8 |
| Altersunterschied | 43,5 | 45,2 | 57,5 |
| Schlaf | 47,7 | 40,5 | 33,8 |
| Größe-Männer | 82,2 | 54,5 | 53,3 |
| Größe-Frauen | 43,3 | 45,6 | 47,8 |

**Tabelle 6.10: Mittelwertdifferenz zwischen Gruppe 1 und Gruppe 2**

Die 7-stufige Ratingskala scheint insbesondere bei den Merkmalen *Altersunterschied* und *Größe-Frauen* gut zu diskriminieren, da hier die Mittelwertdifferenz am höchsten ist.

Die 5-stufige Skala dominiert bei den Merkmalen *Vorlesung*, *Jobben* und *Entfernung*.

Die größte Mittelwertdifferenz liefert nahezu immer die 3-stufige Ratingskala. Dies gilt für die Merkmale *Lernstunden*, *Einkommen*, *Geld*, *Universität*, *Ehrenamt*, *Hobbys*, *Miete*, *Wohnfläche, Geschwister, Schlaf* und *Größe-Männer*.

Auf Basis der Diskriminierungsfähigkeit ist die 3-stufige Skala somit besser als die anderen Skalen. Die 5- und 7-stufigen Skalen sind bei diesem Kriterium nur für sehr wenige Merkmale am besten geeignet.

PRESTON und COLMAN (2000) verwenden bei ihrer Analyse die Teststatistik des Zweistichproben-t-Tests als Vergleichsgröße. Wie bereits in Kapitel 5.3.3 geschildert (siehe Seite 150ff) bevorzugt diese Vorgehensweise Ratingskalen mit vielen Antwortkategorien. Die 3-stufige Skala mit ihren Ausprägungen 100, 50 und 0 weist nämlich meist eine höhere Varianz auf als die 7-stufige Skala mit den Ausprägungen 100, 83, 67, 50, 33, 17 und 0.

Besonders kritisch ist die Interpretation der Teststatistik in der hier vorgestellten Analyse, weil die Stichprobengrößen, also die Anzahl der Objekte in Gruppe 1 und Gruppe 2, bei der Berechnung der Teststatistik einfließen. Durch das gewählte Untersuchungsdesign und die Vorgehensweise bei der Bestimmung der Gruppen ist nicht sichergestellt, dass für die verschiedenen Ratingskalen diese Stichprobengröße identisch ist. Daher soll auf die Bewertung der Teststatistik verzichtet werden.

### 6.2.1.3 Zusammenfassung der Ergebnisse

Die verschiedenen Ratingskalen wurden mit drei verschiedenen Kriterien bewertet:

- Kriteriumsvalidität gemessen mit Kendalls τ-b
- Kriteriumsvalidität gemessen mit dem Rangkorrelationskoeffizienten von Spearman
- Diskriminierungsfähigkeit

Insgesamt kann festgehalten werden, dass die 7-stufige Ratingskala am schlechtesten abschneidet. Um die Vorzüge der einzelnen Ratingskalen zu verdeutlichen, soll im Folgenden getrennt für die verschiedenen Skalen

angegeben werden, bei welchem Kriterium und welcher Frage sie die beste Alternative darstellen.

In Tabelle 6.11 ist zunächst angegeben, wann die *Ratingskala* mit *7 Kategorien* die beste Alternative ist (mit X gekennzeichnet).

| | Kendalls τ-b | Spearman | Mittelwert-differenz |
|---|---|---|---|
| Vorlesung | | | |
| Lernstunden | | | |
| Jobben | | | |
| Einkommen | | | |
| Geld | | | |
| Universität | | | |
| Ehrenamt | | | |
| Entfernung | | | |
| Hobbys | | | |
| Miete | | | |
| Wohnfläche | | | |
| Geschwister | X | X | |
| Altersunterschied | X | X | X |
| Schlaf | | | |
| Größe-Männer | | X | |
| Größe-Frauen | | | X |

**Tabelle 6.11: 7-stufige Ratingskala als bestes Messinstrument**

Die 7-stufige Ratingskala ist für die Fragestellungen *Geschwister* und *Alters-unterschied* am besten geeignet. Beim Merkmal *Geschwister* reicht die Bandbreite der Antworten, bis auf sehr wenige Ausreißer, von 0 bis 6 Geschwistern. Dies führt unter Umständen zu einer Homogenisierung des Abbildungsprozesses. Ein ähnliches Phänomen tritt beim *Altersunterschied* auf. Auch hier sind Werte zwischen 0 und 6 Jahren die Regel. Ausreißer nach oben sind zwar etwas häufiger als beim Merkmal *Geschwister*, was sich in einer insgesamt niedrigeren Korrelation und geringeren Diskriminierungsfähigkeit niederschlägt, dennoch führt die relativ überschaubare Zahl der möglichen Antworten anscheinend zu einer Homogenisierung des Mappingprozesses. Dies

könnte ein Grund sein, weshalb gerade bei diesen Themenstellungen die 7-stufige Ratingskala zu bevorzugen ist.

In Tabelle 6.12 wird deutlich, wann die *5-stufige Ratingskala* zu bevorzugen ist:

| | Kendalls τ-b | Spearman | Mittelwert-differenz |
|---|---|---|---|
| Vorlesung | X | X | X |
| Lernstunden | X | X | |
| Jobben | X | X | X |
| Einkommen | X | X | |
| Geld | X | X | |
| Universität | | | |
| Ehrenamt | | | |
| Entfernung | X | X | X |
| Hobbys | | | |
| Miete | X | X | |
| Wohnfläche | | | |
| Geschwister | | | |
| Altersunterschied | | | |
| Schlaf | X | X | |
| Größe-Männer | X | | |
| Größe-Frauen | X | X | |

**Tabelle 6.12: 5-stufige Ratingskala als bestes Messinstrument**

Bei Verwendung der korrelationsbasierten Kriterien ist die 5-stufige Ratingskala meist am besten geeignet. Bei der Diskriminierungsfähigkeit dominiert die 5-stufige Skala dagegen nur bei den Themenstellungen *Vorlesung, Jobben* und *Entfernung,* da bei diesem Kriterium die 3-stufige Skala häufig besser ist. Allerdings sind die Korrelationsaussagen als Messung der Validität von größerem Gewicht, als die Diskriminierungsfähigkeit. Insofern kann festgehalten werden, dass die 5-stufige Ratingskala bei der Mehrzahl der Fragestellungen das beste Messinstrument darstellt.

Die *3-stufige Ratingskala* schneidet bei den Merkmalen *Universität, Ehrenamt, Hobbys* und *Wohnfläche* am besten ab, wie Tabelle 6.13 zeigt:

| | Kendalls τ-b | Spearman | Mittelwert-differenz |
|---|---|---|---|
| Vorlesung | | | |
| Lernstunden | | | X |
| Jobben | | | |
| Einkommen | | | X |
| Geld | | | X |
| Universität | X | X | X |
| Ehrenamt | X | X | X |
| Entfernung | | | |
| Hobbys | X | X | X |
| Miete | | | X |
| Wohnfläche | X | X | X |
| Geschwister | | | X |
| Altersunterschied | | | |
| Schlaf | | | X |
| Größe-Männer | | | X |
| Größe-Frauen | | | |

**Tabelle 6.13: 3-stufige Ratingskala als bestes Messinstrument**

Die Merkmale *Universität* und *Ehrenamt* weisen die Gemeinsamkeit auf, dass sehr viele Personen beim externen Kriterium 0 angeben. Somit liegen für diese beiden Merkmale sehr viele Bindungen beim externen Kriterium vor, welche hervorragend durch eine 3-stufige Skala abgebildet werden können. Das gute Abschneiden der Ratingskala mit 3 Antwortkategorien ist für diese Merkmale insofern nicht überraschend. Dass die Ratingskala jedoch auch bei den Merkmalen *Hobbys* und *Wohnfläche* am besten geeignet ist ein weiteres Indiz, dass der Abbildungsprozess bereits bei einer 5-stufigen Skala problematisch sein kann.

Bis auf wenige durch die Merkmalsbesonderheiten erklärbare Ausnahmen ist die 7-stufige Skala nicht geeignet, die Information der metrischen Merkmale abzubilden. Beim Vergleich der 3- und 5-stufigen Skala zeichnet sich eine Tendenz zu Gunsten der 5-stufigen Skala ab.

Eine mögliche Erklärung für diese Resultate liefern inhomogene Abbildungsprozesse, die insbesondere bei Ratingskalen mit vielen Antwortkategorien zum tragen kommen.

## 6.2.2 Existenz inhomogener Abbildungsprozesse

Entsprechend **Zielsetzung 2** soll in dieser Studie herausgearbeitet werden, dass inhomogene Mappingprozesse der Probanden für unpräzise Messungen verantwortlich sind. Dazu wird folgende Vorgehensweise gewählt:

- Einen ersten Hinweis auf die Inhomogenität der Abbildungsprozesse liefert die graphische Veranschaulichung der Ergebnisse mit Hilfe von Boxplots (Kapitel 6.2.2.1).

- Boxplots liefern Anhaltspunkte für die Existenz unterschiedlicher Abbildungsprozesse. Neben diesen Mappingprozessen können aber auch das Abrufen und die Beurteilung vorhandener Information für die beobachteten Phänomene verantwortlich sein. Kapitel 6.2.2.2 greift deshalb auf die Erhebung der Einschätzung einer Durchschnittsperson ($z_i$) zurück, um aufzuzeigen, dass auch bei nahezu identischen Einschätzungen der Durchschnittsperson Beantwortungseffekte zu beobachten sind. Da diese dann nicht auf das Abrufen und die Beurteilung relevanter Information zurückgeführt werden können, sind sie ein weiterer Beleg für die Existenz inhomogener Abbildungsprozesse.

### 6.2.2.1 Graphische Veranschaulichung der Existenz inhomogener Abbildungsprozesse

Wie Boxplots bei der Beurteilung von Messinstrumenten eingesetzt werden können, ist in Kapitel 3.4.4.3 (Seite 94ff.) ausführlich geschildert. Die Einteilung der verschiedenen Gruppen basiert auf den Ratingskalen ($x_j$). Mit Hilfe der Verteilung des betreffenden externen Kriteriums ($y_j$) werden anschließend die Boxplots getrennt für jede Gruppe bestimmt.

Die folgenden Ausführungen sollen sich auf das Merkmal *Schlaf* beschränken, Boxplots für weitere Themenstellungen führen zu ähnlichen Ergebnissen. Bei der

Erstellung der Boxplots erfolgt eine Einteilung der Personen auf Basis der Antwort zur (Rating-)Frage: ´Ich benötige viel Schlaf, um mich fit zu fühlen´ ($x_{14}$). Bei der 3-stufigen Skala werden somit 3 Gruppen gebildet:

- Personen, die mit ´stimme überhaupt nicht zu´ antworten (Gruppe 1)
- Personen, die mit ´teils teils´ antworten (Gruppe 2)
- Personen, die mit ´stimme voll und ganz zu´ antworten (Gruppe 3)

Bei der 5- bzw. 7-stufigen Skala werden analog 5 bzw. 7 Gruppen gebildet. Anschließend erfolgt für jede Gruppe getrennt die Berechnung der Boxplots auf Basis der metrischen Variable ($y_{14}$), also der Antwort auf die Frage: ´Wie viele Stunden Schlaf benötigen Sie, um sich fit zu fühlen?´

In Abbildung 6.6 sind die so gebildeten Boxplots für drei Antwortkategorien dargestellt:

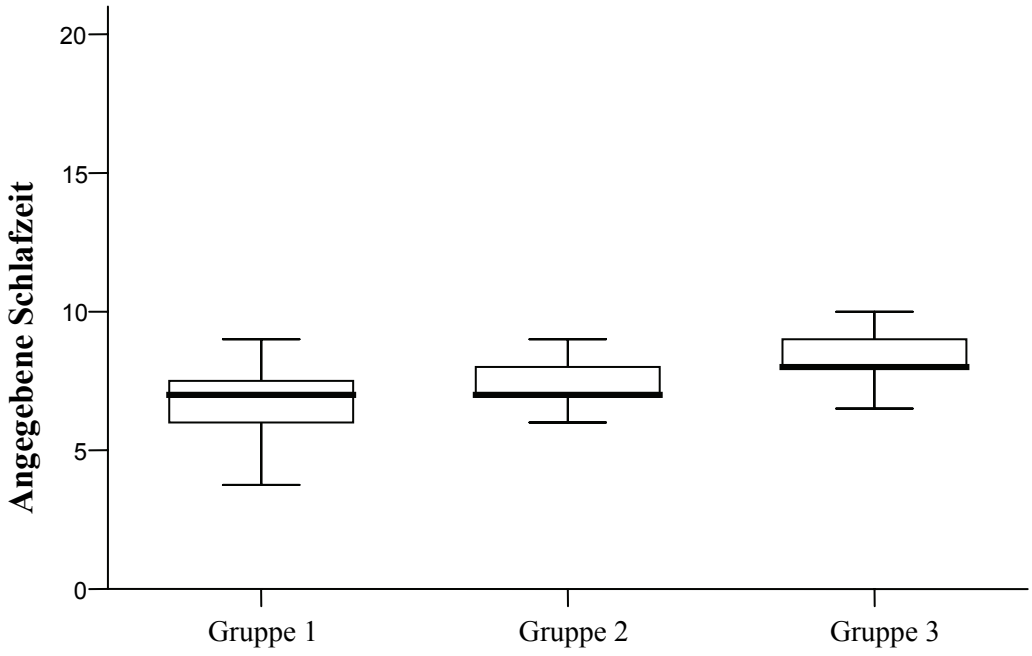


**Abbildung 6.6: Boxplot für drei Antwortkategorien**

Abbildung 6.6 zeigt grundsätzlich die erwarteten Tendenzen. So liegen alle relevanten Quantile von Personengruppe 3 über denen der anderen beiden

Gruppen. Allerdings zeigt ein Vergleich der Boxplots auch, dass es zwischen den einzelnen Gruppen zu deutlichen Überschneidungen kommt. Ein nicht unerheblicher Teil der Probanden aus Gruppe 1 benötigt mehr Schlaf als Personen aus Gruppe 2. So liegt z.B. der Median der Gruppe 1 über dem 25%-Quantil der Gruppe 2.

Die Überschneidungen nehmen bei der 5-stufigen Ratingskala weiter zu (siehe Abbildung 6.7).

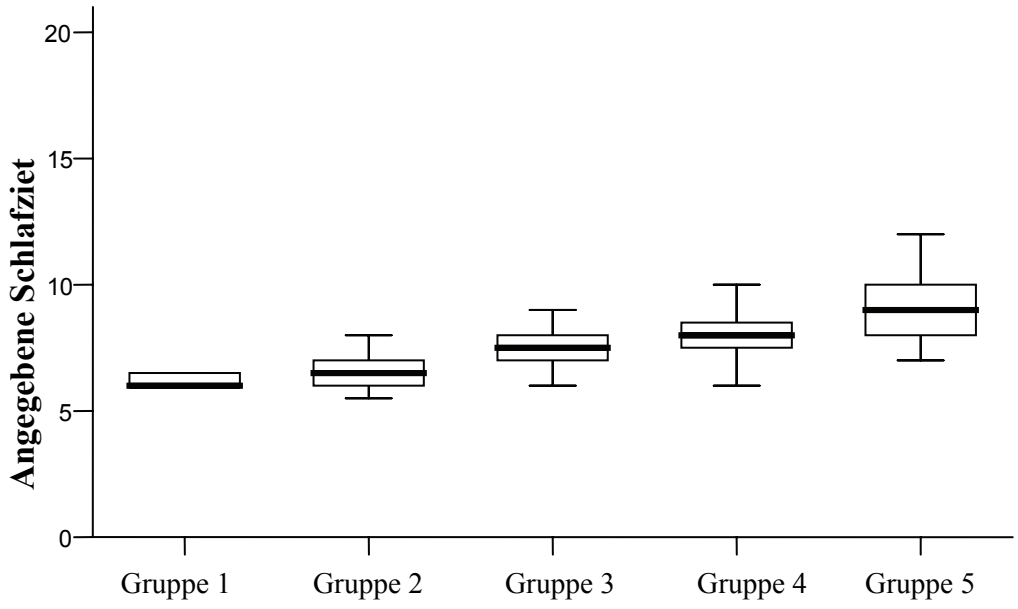


**Abbildung 6.7: Boxplot für fünf Antwortkategorien**

Bei 7 Antwortkategorien umfasst Gruppe 1 lediglich zwei Personen, sie sollte deshalb nur mit Einschränkungen interpretiert werden. Dennoch verdeutlichen die Boxplots für diese Ratingskala noch stärker, wie groß das Problem inhomogener Abbildungsprozesse ist (siehe Abbildung 6.8):

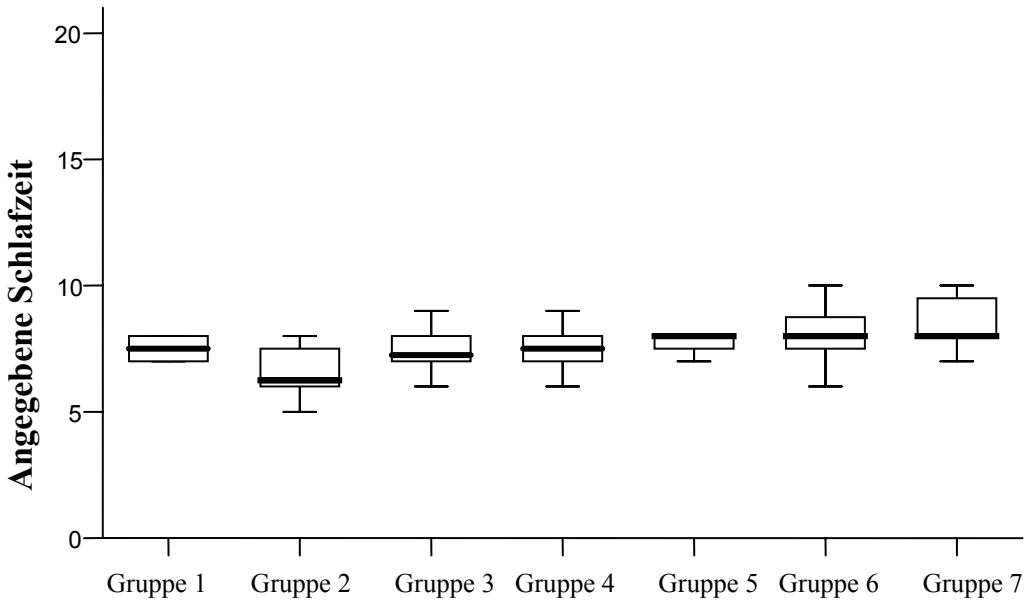


**Abbildung 6.8: Boxplot für sieben Antwortkategorien**

Diese einführenden Erläuterungen machen deutlich, dass inhomogene Mappingprozesse die Ergebnisse einer Ratingskala verzerren können. Für diese Verzerrung könnte jedoch auch die inhomogene Informationsbeurteilung verantwortlich sein. Daher soll im Folgenden der Versuch unternommen werden, eine einheitliche Informationsbeurteilung der Probanden sicherzustellen.

### 6.2.2.2 Berücksichtigung inhomogener Informationsbeurteilung

Bei den abgefragten Ratingskalen müssen die Versuchspersonen eine Einschätzung der eigenen Person im Vergleich zu einer durchschnittlichen Person einer Vergleichsgruppe vornehmen. Daher ist es für die Probanden erforderlich, zunächst durch Abrufen und Beurteilung vorhandener Information zu einem Ergebnis für eine durchschnittliche Person zu gelangen, beispielsweise wie viel Miete ein durchschnittlicher Student bezahlen muss. Dafür kann der studentische Freundeskreis einen Anhaltspunkt liefern, aber auch Medienberichte oder Erzählungen aus der Verwandtschaft. Dabei kommen Probanden selbst dann nicht zu identischen Ergebnissen, wenn sie dieselbe Strategie verwenden, etwa die durchschnittliche Miete im Freundeskreis, da diese sich sehr stark unterscheiden kann.

Die Relevanz differierender Informationsbeurteilung wird auch dadurch ersichtlich, dass insbesondere bei jenen Themenstellungen, die eine homogene Einschätzung einer Durchschnittsperson realistisch erscheinen lassen (*Größe, Geschwister, Altersunterschied*) die Korrelationen zwischen den Ratingskalen ($x_j$) und externen Kriterien ($y_j$) besonders hoch sind (vgl. Tabelle 6.5 und 6.6).

Deshalb ist im Fragebogen die Einschätzung eines durchschnittlichen Studenten explizit nochmals erhoben ($z_j$). Mit Hilfe dieser Angabe können die Studenten in zwei Gruppen A und B eingeteilt werden. Dazu werden zunächst 25%- und 75%-Quantil der ($z_j$) bestimmt.

Die Personengruppe A setzt sich aus solchen Probanden zusammen, deren Einschätzung einer Durchschnittsperson kleiner oder gleich dem 25%-Quantil von $(z_j)$ ist. Personengruppe B besteht dagegen aus jenen Studenten, deren Einschätzung größer oder gleich dem 75%-Quantil von $(z_j)$ ist. Beide Personengruppen für sich betrachtet sind somit hinsichtlich der Einschätzung einer Durchschnittsperson sehr homogen. Daher sollte die Korrelation zwischen Ratingskala und externem Kriterium innerhalb der Personengruppen sehr hoch sein.

Bei den Merkmalen *Universität* und *Ehrenamt* geben über die Hälfte der Probanden den Wert 0 für $z_6$ und $z_7$ ab. Personengruppe A würde deshalb aus mehr als der Hälfte der Probanden bestehen. Deshalb werden diese beiden Themenstellungen bei den folgenden Auswertungen nicht berücksichtigt.

In Tabelle 6.14 sind die Korrelationen (Kendalls τ-b) der 3-stufigen Ratingskalen $(x_j)$ mit den externen Kriterien $(y_j)$ für Personengruppe A und B angegeben. Zusätzlich ist in der ersten Spalte die ursprüngliche Korrelation bei Berücksichtigung aller Personen dargestellt (vgl. Tabelle 6.5):

| | Alle Personen | Personengruppe A | Personengruppe B |
|---|---|---|---|
| Vorlesung ($r_{x_1y_1}^{\tau-b}$) | 0,387 | 0,411 | 0,521 |
| Lernstunden ($r_{x_2y_2}^{\tau\text{-}b}$) | 0,279 | 0,276 | 0,356 |
| Jobben ($r_{x_3y_3}^{\tau\text{-}b}$) | 0,566 | 0,551 | 0,569 |
| Einkommen ($r_{x_4y_4}^{\tau\text{-}b}$) | 0,388 | 0,444 | 0,545 |
| Geld ($r_{x_5y_5}^{\tau\text{-}b}$) | 0,355 | 0,386 | 0,478 |
| Entfernung ($r_{x_8y_8}^{\tau\text{-}b}$) | 0,382 | 0,350 | 0,416 |
| Hobbys ($r_{x_9y_9}^{\tau\text{-}b}$) | 0,447 | 0,491 | 0,477 |
| Miete ($r_{x_{10}y_{10}}^{\tau\text{-}b}$) | 0,523 | 0,631 | 0,547 |
| Wohnfläche ($r_{x_{11}y_{11}}^{\tau\text{-}b}$) | 0,500 | 0,504 | 0,578 |

**Tabelle 6.14: Korrelation der 3-stufigen Skala für Personengruppe A und B**

Mit wenigen Ausnahmen steigt die Korrelation, wenn sie innerhalb der homogenen Personengruppen A und B gebildet wird. Ein ähnliches Ergebnis liefert ebenfalls die 5-stufige Ratingskala, die in Tabelle 6.15 dargestellt ist.

| | Alle Personen | Personengruppe A | Personengruppe B |
|---|---|---|---|
| Vorlesung ($r_{x_1y_1}^{\tau-b}$) | 0,561 | 0,596 | 0,581 |
| Lernstunden ($r_{x_2y_2}^{\tau\text{-}b}$) | 0,297 | 0,415 | 0,422 |
| Jobben ($r_{x_3y_3}^{\tau\text{-}b}$) | 0,632 | 0,710 | 0,712 |
| Einkommen ($r_{x_4y_4}^{\tau\text{-}b}$) | 0,392 | 0,426 | 0,463 |
| Geld ($r_{x_5y_5}^{\tau\text{-}b}$) | 0,418 | 0,452 | 0,479 |
| Entfernung ($r_{x_8y_8}^{\tau\text{-}b}$) | 0,483 | 0,384 | 0,481 |
| Hobbys ($r_{x_9y_9}^{\tau\text{-}b}$) | 0,361 | 0,418 | 0,395 |
| Miete ($r_{x_{10}y_{10}}^{\tau\text{-}b}$) | 0,665 | 0,707 | 0,724 |
| Wohnfläche ($r_{x_{11}y_{11}}^{\tau\text{-}b}$) | 0,406 | 0,475 | 0,556 |

**Tabelle 6.15: Korrelation der 5-stufigen Skala für Personengruppe A und B**

In Tabelle 6.16 sind schließlich die Ergebnisse für die 7-stufige Skala abgebildet, die nochmals die Ergebnisse der 3- und 5-stufige Skala bestätigen.

| | Alle Personen | Personengruppe A | Personengruppe B |
|---|---|---|---|
| Vorlesung ( $r_{x_1 y_1}^{\tau - b}$ ) | 0,420 | 0,499 | 0,676 |
| Lernstunden ( $r_{x_2 y_2}^{\tau - b}$ ) | 0,160 | 0,154 | 0,211 |
| Jobben ( $r_{x_3 y_3}^{\tau - b}$ ) | 0,554 | 0,619 | 0,581 |
| Einkommen ( $r_{x_4 y_4}^{\tau - b}$ ) | 0,255 | 0,246 | 0,332 |
| Geld ( $r_{x_5 y_5}^{\tau - b}$ ) | 0,300 | 0,576 | 0,491 |
| Entfernung ( $r_{x_8 y_8}^{\tau - b}$ ) | 0,328 | 0,293 | 0,418 |
| Hobbys ( $r_{x_9 y_9}^{\tau - b}$ ) | 0,385 | 0,670 | 0,492 |
| Miete ( $r_{x_{10} y_{10}}^{\tau - b}$ ) | 0,528 | 0,527 | 0,579 |
| Wohnfläche ( $r_{x_{11} y_{11}}^{\tau - b}$ ) | 0,448 | 0,509 | 0,536 |

**Tabelle 6.16: Korrelation der 7-stufigen Skala für Personengruppe A und B**

Der Rangkorrelationskoeffizient nach Spearman führt zu vergleichbaren Ergebnissen, daher soll er hier nicht aufgeführt werden.

Mit Ausnahme des Merkmals *Entfernung* steigt die Korrelation durch die Bildung homogener Subgruppen (vgl. Tabelle 6.14, 6.15 und 6.16). Die Qualität der Befragungsergebnisse steigt also mit zunehmender Homogenität der Informationsbeurteilung. Dennoch sind auch innerhalb der Personengruppen A und B die Korrelationen zum Teil sehr niedrig. Dies ist ein weiteres Indiz für die Existenz inhomogener Abbildungsprozesse. Denn diese liefern eine Erklärung, weshalb trotz homogener Informationsbeurteilung die Messung mit einer Ratingskala Beantwortungseffekte hervorruft.

## 6.2.3 Identifikation unterschiedlicher Mappingstrategien

Falls unterschiedliche Mappingstrategien existieren, ist die Frage nahe liegend, ob die Messung mit Hilfe der Ratingskala durch die Identifikation dieser Abbildungsprozesse verbessert werden kann (**Zielsetzung 3**). Die grundlegende Idee besteht demzufolge darin, das Antwortverhalten einzelner Personen zu identifizieren. Wenn eine Person beispielsweise bei den Ratingskalen ($x_j$)

bevorzugt Extremkategorien ankreuzt, so sollte dies bei der Zuweisung von Skalenwerten zu den einzelnen Antwortkategorien berücksichtigt werden.

Inhomogene Mappingstrategien führen vor allem bei Ratingskalen mit vielen Antwortkategorien zu Verzerrungen. Daher werden bei den folgenden Ausführungen ausschließlich 5- und 7-stufige Ratingskalen betrachtet. Ziel ist es, die Skalenwerte der Ratingskalen ($x_j$) unter Berücksichtigung des individuellen Abbildungsprozesses so zu verändern, dass die Korrelation mit dem externen Kriterium ($y_j$) erhöht wird.

Wie diese Veränderung der Skalenwerte unter Berücksichtigung des individuellen Abbildungsprozesses vorgenommen wird, ist im Folgenden diskutiert.

- In Kapitel 6.2.3.1 wird dazu die gesamte Information aller erhobenen Ratingskalen und externen Kriterien des Fragebogens genutzt. Hierbei stellt sich die Frage nach der Umsetzbarkeit in der Praxis.

- Deshalb ist für reale Befragungssituationen die Überlegung sinnvoll, ob die Ratingskalen verbessert werden können, ohne ein externes Kriterium zu erheben (Kapitel 6.2.3.2).

- Kapitel 6.2.3.3 versucht schließlich Erkenntnisse über das individuelle Antwortverhalten auf Basis eines einzelnen repräsentativen externen Kriteriums zu gewinnen.

### 6.2.3.1 Identifikation der Mappingstrategie mit allen erhobenen Daten

Die Merkmale werden sowohl mit Hilfe des metrischen externen Kriteriums ($y_j$), als auch mit der Ratingskala ($x_j$) gemessen. Bildet man für die metrischen Merkmale ($y_j$) eine Rangfolge der Probanden, um damit die Personen in fünf bzw. sieben etwa gleich große Gruppen einzuteilen, gewinnt man einen Eindruck, welche Kategorie der Proband auf der 5- bzw. 7-stufigen Ratingskala hätte ankreuzen müssen ($x_j^{erwartet}$). Das Merkmal *Größe* wird bei den folgenden Ausführungen nicht berücksichtigt, da nach einer Trennung der Geschlechter die Personenzahl pro Gruppe sehr gering ist.

Abbildung 6.9 veranschaulicht dies exemplarisch für 10 Probanden, die auf Basis des externen Kriteriums in 5 Gruppen eingeteilt werden:

| | $y_1 \rightarrow x_1^{erwartet}$ | | | $y_{14} \rightarrow x_{14}^{erwartet}$ | |
|---|---|---|---|---|---|
| Person | $y_1$ | $x_1^{erwartet}$ | … | $y_{14}$ | $x_{14}^{erwartet}$ |
| 1 | 0 | 1 | … | 6 | 1 |
| 2 | 1 | 1 | … | 6,5 | 1 |
| 3 | 3 | 2 | … | 7 | 2 |
| 4 | 4 | 2 | … | 7 | 2 |
| 5 | 5 | 3 | … | 7,5 | 3 |
| 6 | 5 | 3 | … | 8 | 3 |
| 7 | 6 | 4 | … | 8,5 | 4 |
| 8 | 7 | 4 | … | 8,5 | 4 |
| 9 | 10 | 5 | … | 9 | 5 |
| 10 | 20 | 5 | … | 9 | 5 |

**Abbildung 6.9: Bestimmung von $x_j^{erwartet}$ mit Hilfe des externen Kriteriums**

Anschließend können zwei Kennzahlen berechnet werden:

- Es wird für jeden Studenten gezählt, wie häufig er bei der Ratingskala ($x_j$) eine Antwortkategorie am Rand wählt, obwohl er auf Basis der Gruppeneinteilung des metrischen Merkmals ($y_j$) einer der mittleren Gruppen angehört. Bei der 5-stufigen Ratingskala wird also für jede Versuchsperson gezählt, wie häufig gleichzeitig

  - $x_j$ den Wert 1 oder 5 annimmt und
  - $x_j^{erwartet}$ die Werte 2 bis 4

  Ein hoher Wert bei diesem Kriterium (*Rand*) spricht dafür, dass der betreffende Student zu Randkategorien tendiert.

- Außerdem wird ausgewertet, wie häufig ein Proband eine Antwortkategorie in der Mitte präferiert, obwohl er sich auf Basis des externen Kriteriums in einer Randgruppe befindet. Bei der 5-stufigen Ratingskala wird bei dieser Kennzahl somit für jeden Probanden gezählt, wie häufig gleichzeitig

  - $x_j$ die Werte 2 bis 4 annimmt und

- $x_j^{erwartet}$ die Werte 1 oder 5

Ein hoher Wert bei dieser Kennzahl (*Mitte*) ist ein Anzeichen dafür, dass die Person eher zu Antwortkategorien in der Mitte neigt.

Der Skalenwert bei der Ratingskala ($x_j$) kann nun mit Hilfe der Kennzahlen *Rand* und *Mitte* modifiziert werden. (6.3) zeigt diese Transformation für die 5-stufige Ratingskala:

$$x_{ij}^{neu} = \begin{cases} x_{ij} + \frac{R_i}{14} & \text{für } x_{ij} = 1 \\ x_{ij} - \frac{M_i}{14} & \text{für } x_{ij} = 2 \\ x_{ij} & \text{für } x_{ij} = 3 \\ x_{ij} + \frac{M_i}{14} & \text{für } x_{ij} = 4 \\ x_{ij} - \frac{R_i}{14} & \text{für } x_{ij} = 5 \end{cases} \tag{6.3}$$

$x_{ij}^{neu}$: *Neuer (transformierter) Skalenwert von Person i bei Merkmal* $x_j$

$x_{ij}$: *Ursprünglicher Skalenwert von Person i bei Merkmal* $x_j$

$R_i$: *Ausprägung der Kennzahl Rand bei Person i*

$M_i$: *Ausprägung der Kennzahl Mitte bei Person i*

Die Skalenwerte an den Rändern der ursprünglichen Ratingskala (1 und 5) müssen in Richtung der Mitte verschoben werden, wenn ein Proband einen hohen Wert bei *Rand* hat. Dafür wird zunächst die Kennzahl *Rand* durch die Anzahl der Merkmale (14) dividiert. Dieser Quotient wird beim Skalenwert 1 addiert und beim Skalenwert 5 subtrahiert. Die Division durch 14 bewirkt, dass jeder Skalenwert um maximal 1 verschoben wird. So bleibt die grundlegende Tendenz der Antwort des Probanden erhalten.

Mit Hilfe der Kennzahl *Mitte* werden bei der 5-stufigen Skala die Antworten 2 und 4 abgeändert. Bei Personen mit einem hohen Wert bei *Mitte* erfolgt eine Anpassung in Richtung des Rands der Ratingskala. Dazu wird die Kennzahl *Mitte* ebenfalls durch 14 dividiert, um die prinzipielle Struktur der Antworten eines Probanden zu erhalten. Der Quotient wird vom Skalenwert 2 abgezogen und zum Skalenwert 4 addiert.

Für die 7-stufige Skala gilt bei identischer Notation entsprechend:

$$x_{ij}^{neu} = \begin{cases} x_{ij} + \frac{R_i}{14} & \text{für } x_{ij} = 1 \\ x_{ij} - \frac{M_i}{14} & \text{für } x_{ij} = 2,3 \\ x_{ij} & \text{für } x_{ij} = 4 \\ x_{ij} + \frac{M_i}{14} & \text{für } x_{ij} = 5,6 \\ x_{ij} - \frac{R_i}{14} & \text{für } x_{ij} = 7 \end{cases} \tag{6.4}$$

Für die 7-stufige Skala wäre es auch denkbar, die Skalenwerte 2 und 6 nicht zu verändern, da diese nicht eindeutig den mittleren Kategorien zugeordnet werden können. Die empirische Untersuchung ergibt für diesen Fall keine nennenswerten Unterschiede der Ergebnisse, weshalb auf eine detaillierte Beschreibung der Ergebnisse dieser Vorgehensweise verzichtet wird.

Eine Sonderrolle nimmt die mittlere Kategorie ein. Gerade diese Kategorie wird besonders häufig von jenen Personen gewählt, die eine Extremkategorie vermeiden. Allerdings kann für diese Kategorie nicht beantwortet werden, in welche Richtung die Antwort angepasst werden soll. Daher bleibt diese Kategorie bei diesem Konzept unverändert.

In Tabelle 6.17 ist die Korrelation zwischen der 5-stufigen Skala und dem jeweiligen externen Kriterium, gemessen mit Kendalls τ-b, angegeben. In der mittleren Spalte sind die ursprünglichen Korrelationen aufgelistet (vgl. Tabelle 6.5). Die rechte Spalte stellt die Korrelation der nach (6.3) neu berechneten Werte mit dem externen Kriterium dar.

| | $r_{x_j y_j}^{\tau\text{-b}}$ | $r_{x_j^{neu} y_j}^{\tau\text{-b}}$ |
|---|---|---|
| Vorlesung | 0,561 | 0,616 |
| Lernstunden | 0,297 | 0,344 |
| Jobben | 0,632 | 0,641 |
| Einkommen | 0,392 | 0,437 |
| Geld | 0,418 | 0,448 |
| Univesität | 0,342 | *0,319* |
| Ehrenamt | 0,540 | 0,552 |
| Entfernung | 0,483 | 0,547 |
| Hobbys | 0,361 | 0,402 |
| Miete | 0,665 | 0,704 |
| Wohnfläche | 0,406 | 0,468 |
| Geschwister | 0,707 | *0,653* |
| Altersunterschied | 0,502 | *0,458* |
| Schlaf | 0,510 | 0,545 |

**Tabelle 6.17: Korrelation (Kendalls τ-b) zwischen $x_j^{neu}$ und $y_j$**

Die Korrelation zwischen den transformierten Skalenwerten und dem metrischen Merkmal steigt im Vergleich zur Korrelation der nicht veränderten Ratingskalen mit dem metrischen Merkmal.

Lediglich eine minimale Erhöhung der Korrelation ergibt sich bei den Merkmalen *Jobben* und *Ehrenamt*. Bei den Merkmalen *Universität*, *Geschwister* und *Altersunterschied* sinkt die Korrelation sogar. Bei all diesen Merkmalen antwortet ein relativ großer Teil der Personen beim metrischen Merkmal mit 0. Bei einer näheren Betrachtung der Merkmale fällt auf, dass gerade diese Personen relativ einheitlich auf der Ratingskala den kleinsten Wert ankreuzen. Durch die Transformation dieser Skalenwerte wird eine Rangfolge in diese Studenten gebracht, die so nicht beim metrischen Merkmal vorzufinden ist. Daher führt die Transformation der Skalenwerte für diese Merkmale zu keiner Verbesserung oder sogar zu einer Verschlechterung.

Für die 7-stufige Skala sind vergleichbare Ergebnisse erzielbar. Ähnliche Tendenzen können auch mit dem Rangkorrelationskoeffizienten aufgezeigt werden. Sie sollen an dieser Stelle jedoch nicht näher ausgeführt werden, da die Umsetzbarkeit dieses Ansatzes in der Praxis bezweifelt werden muss.

Die Berechnung der Kriterien *Rand* und *Mitte* kann nur erfolgen, weil in der empirischen Studie das metrische externe Kriterium ($y_i$) gleichzeitig erhoben wurde. In der Realität misst man mit Hilfe der Ratingskala jedoch meist Sachverhalte, für die eine Messung des externen Kriteriums unmöglich ist. Insofern liefern die Ergebnisse lediglich einen Anhaltspunkt, ob die Identifikation unterschiedlicher Mappingstrategien theoretisch die Ergebnisse verbessern kann. Für die Anwendung in der Praxis ist dieses Konzept dagegen meist nicht zu gebrauchen. Diesen Nachteil versuchen die folgenden Methoden zu umgehen.

### 6.2.3.2 Identifikation der Mappingstrategie ohne externes Kriterium

Im Rahmen dieses Abschnitts wird erneut auf die Kennzahlen *Rand* und *Mitte* zurückgegriffen. Allerdings sollen die beiden Kennzahlen jetzt, wie in der Praxis eher möglich, ohne die Berücksichtigung des externen Kriteriums ermittelt werden. Dazu wird für jeden Probanden die Anzahl der Extremantworten bei den einzelnen Ratingskalen ($x_j$), also die Kategorien 1 und 7 bei der 7-stufigen Ratingskala, gezählt. Diese Anzahl entspricht dem Kriterium *Rand*. Die Kennzahl *Mitte* entspricht analog der Anzahl der Antworten mit einer der mittleren Kategorien (3 bis 5 bei der 7-stufigen Skala).

Diese Vorgehensweise verändert aber auch die Skalenwerte für jene Personen, die eine Randkategorie deshalb wählen, weil sie tatsächlich eine extreme metrische Ausprägung haben. Diesen Probanden wird bei dieser Methode nämlich fälschlicherweise unterstellt, dass sie zu Extremantworten neigen, da keine Vergleichsmöglichkeiten mit einem externen Kriterium exisitieren. Neben dieser Berücksichtigung der Kategorien *Rand* und *Mitte*, mit der versucht wird, das Antwortverhalten einer Versuchsperson in Bezug auf die Neigung zu Extremantworten zu ermitteln, ist darüber hinaus noch eine weitere Transformation vorstellbar: Die generelle Tendenz der Probanden zur *Zustimmung* oder *Ablehnung*.

Unter Umständen neigt ein Teil der Probanden dazu, bei den Ratingfragen ($x_j$) eher zuzustimmen, während ein anderer Teil zur Ablehnung tendiert. Wenn ein Proband zur Ablehnung einer Frage tendiert, sind seine Skalenwerte zu niedrig angesetzt. Er gibt zwar die Einschätzung ´stimme eher nicht zu´ (Skalenwert 3) ab, diese ist aber vergleichbar mit der Aussage ´teils teils´ (Skalenwert 4) einer

Person, die keine Antworttendenzen vorweist. Dagegen sind die Skalenwerte zu hoch angesetzt, wenn ein Proband generell zur Zustimmung neigt. Es erscheint deshalb sinnvoll, die Anzahl der Antworten, die eine *Ablehnung* bzw. *Zustimmung* ausdrücken zu zählen und bei der Transformation der Werte zu berücksichtigen:

$$x_{ij}^{neu} = \begin{cases} x_{ij} + \dfrac{R_i + A_i - Z_i}{14} & \text{für } x_{ij} = 1 \\ x_{ij} - \dfrac{M_i - A_i + Z_i}{14} & \text{für } x_{ij} = 2,3 \\ x_{ij} + \dfrac{A_i - Z_i}{14} & \text{für } x_{ij} = 4 \\ x_{ij} + \dfrac{M_i + A_i - Z_i}{14} & \text{für } x_{ij} = 5,6 \\ x_{ij} - \dfrac{R_i - A_i + Z_i}{14} & \text{für } x_{ij} = 7 \end{cases} \tag{6.5}$$

$x_{ij}^{neu}$: *Neuer (transformierter) Skalenwert von Person i bei Merkmal* $x_j$

$x_{ij}$: *Ursprünglicher Skalenwert von Person i bei Merkmal* $x_j$

$R_i$: *Ausprägung der Kennzahl Rand bei Person i*

$M_i$: *Ausprägung der Kennzahl Mitte bei Person i*

$A_i$: *Anzahl der Merkmale, bei denen der Proband i eine Ablehnung ausdrückt*

$Z_i$: *Anzahl der Merkmale, bei denen der Proband i eine Zustimmung ausdrückt*

In Tabelle 6.18 sind die Korrelationen der 7-stufigen Skala mit dem externen Kriterium nach Kendalls τ-b angegeben. In der mittleren Spalte sind zum Vergleich die ursprünglichen Korrelationen der unveränderten Ratingskalen ($x_j$) mit den externen Kriterien ($y_j$) dargestellt. Die Korrelationen der nach (6.5) transformierten Werte ($x_j^{neu}$) mit den externen Kriterien ($y_j$) befinden sich in der rechten Spalte:

| | $r_{x_j y_j}^{\tau\text{-b}}$ | $r_{x_j^{neu} y_j}^{\tau\text{-b}}$ |
|---|---|---|
| Vorlesung | 0,420 | 0,393 |
| Lernstunden | 0,160 | 0,146 |
| Jobben | 0,554 | 0,520 |
| Einkommen | 0,255 | 0,244 |
| Geld | 0,300 | 0,280 |
| Uni | 0,365 | 0,332 |
| Ehrenamt | 0,521 | 0,494 |
| Entfernung | 0,328 | 0,313 |
| Hobbys | 0,385 | 0,376 |
| Miete | 0,528 | 0,520 |
| Wohnfläche | 0,448 | 0,429 |
| Geschwister | 0,760 | 0,689 |
| Altersunterschied | 0,584 | 0,544 |
| Schlaf | 0,403 | 0,373 |

**Tabelle 6.18: Korrelation zwischen transformierten Werten und externem Kriterium**

Die Korrelation zum externen Kriterium ist bei allen Merkmalen nach der Transformation kleiner. Die Ergebnisse für den Rangkorrelationskoeffizienten sind vergleichbar. Die niedrige Korrelation bedeutet eine Verschlechterung der Ergebnisse, die Transformation ist daher ungeeignet.

Um alle Aspekte bei der Bewertung der Ergebnisse miteinzubeziehen, werden noch folgende Änderungen bei der Berechnung von ($x_j^{neu}$) ergänzt:

- Zusätzliche Berücksichtigung der Kategorien 2 und 6 bei der Kennzahl *Rand*
- Beschränkung auf die Kategorie 4 bei der Kennzahl *Mitte*
- Änderung des Nenners in (6.5)
- Verzicht auf die Kennzahlen *Rand, Mitte*
- Verzicht auf die Kennzahlen *Zustimmung, Ablehnung*

Wie sich zeigt, ziehen diese Ergänzungen jedoch keine Verbesserung der Ergebnisse nach sich.

Das Ziel dieser zweiten Methode ist die Verbesserung der Qualität einer Messung in realen Befragungssituationen. Deshalb wird auf die Erhebung des externen Kriteriums ($y_j$) verzichtet. Es muss jedoch festgehalten werden, dass die Verbesserung der Messergebnisse ohne die Erhebung eines externen Kriteriums unmöglich ist.

### 6.2.3.3 Identifikation der Mappingstrategie mit einem metrischen Merkmal

Die Ergebnisse in Kapitel 6.2.3.2 haben gezeigt, dass die Erhöhung der Qualität einer Ratingskala ohne die Erhebung eines externen Kriteriums nicht erfolgen kann. Wie aber bereits in Kapitel 6.2.3.1 ausgeführt, liegen Messergebnisse für externe Kriterien in realen Befragungssituationen nicht vor. Deshalb soll in diesem Kapitel diskutiert werden, ob es sinnvoll ist, speziell zum Zweck der Verbesserung der Qualität der Messung ein Merkmal mittels Ratingskala und metrisch zu erheben. Wenn diese Vorgehensweise zu einer deutlichen Verbesserung der Ergebnisse führt, scheint der damit einhergehende Anstieg des Erhebungsumfangs gerechtfertigt.

Im Rahmen dieser Methode sollen die Ergebnisse der Ratingskalen ($x_j$) unter Berücksichtigung der metrischen Ausprägung des *Altersunterschieds* der Eltern ($y_{13}$) und des mit der Ratingskala gemessenen *Altersunterschieds* ($x_{13}$) verbessert werden. Die Auswahl des *Altersunterschieds* erfolgt nicht zufällig[3]. Zum einen erscheint dessen Erhebung praktikabel und allgemein übertragbar. Dies gilt z.B. nicht für die Fragestellungen *Vorlesung, Lernstunden, Jobben, Universität, Ehrenamt, Entfernung*, und *Miete.* Diese Fragen können in realen Fragesituationen wahrscheinlich von vielen Probanden nicht beantwortet werden. Zum anderen ist die Frage nach dem Altersunterschied der Eltern keine intime Frage, wie etwa die Fragen *Einkommen*, *Geld* und *Wohnfläche*. Eine einleitende Frage nach den *Schlaf*gewohnheiten, *Hobbys* oder der *Größe* einer Person mutet in den meisten Fragebögen wohl ebenfalls seltsam an. Daher erscheinen einzig die Merkmale *Geschwister* und *Altersunterschied* sinnvoll. Die Ergebnisse für

[3] Die Frage nach dem Altersunterschied ist kritisch, wenn ein Elternteil der Auskunftsperson bereits gestorben ist. Daher empfiehlt es sich, lediglich das Geburtsjahr der Eltern zu ermitteln und den Altersunterschied so zu berechnen.

das Merkmal *Altersunterschied* werden hier vorgestellt, da dessen Verwendung zu etwas besseren Ergebnissen führt als das Merkmal *Geschwister*.

Die Information über das Antwortverhalten bei ($x_{13}$) und ($y_{13}$) kann dazu genutzt werden, das Antwortverhalten der Personen zu identifizieren. Wenn sich die Qualität der Ratingskalen ($x_j$) dadurch deutlich verbessert, scheint die damit einhergehende Erhöhung des Fragebogenumfangs gerechtfertigt.

Zur Transformation der Werte werden erneut die Kennzahlen *Rand* und *Mitte* benötigt. Allerdings wird dazu jetzt lediglich ein Vergleich von $y_{13}$ und $x_{13}$ vorgenommen. Die veränderten Werte ($x_j^{neu}$) berechnen sich dann wie folgt:

$$x_{ij}^{neu} = \begin{cases} x_{ij} + \frac{R_i}{2} & \text{für } x_{ij} = 1 \\ x_{ij} - \frac{M_i}{2} & \text{für } x_{ij} = 2,3 \\ x_{ij} & \text{für } x_{ij} = 4 \\ x_{ij} + \frac{M_i}{2} & \text{für } x_{ij} = 5,6 \\ x_{ij} - \frac{R_i}{2} & \text{für } x_{ij} = 7 \end{cases} \qquad (6.6)$$

$x_{ij}^{neu}$: *Neuer (transformierter) Skalenwert von Person i bei Merkmal* $x_j$

$x_{ij}$: *Ursprünglicher Skalenwert von Person i bei Merkmal* $x_j$

$R_i$: *Ausprägung der Kennzahl Rand bei Person i*

$M_i$: *Ausprägung der Kennzahl Mitte bei Person i*

Wie bereits bei den Methoden in den Kapiteln 6.2.3.1 und 6.2.3.2 wird der Nenner erneut so festgelegt, dass die prinzipielle Antworttendenz eines Probanden erhalten bleibt. Durch den Nenner 2 werden die einzelnen Skalenwerte maximal um 0,5 verschoben.

Die Korrelationen in Tabelle 6.19 zeigen, dass eine Verbesserung der Ergebnisse bei dieser Methode nur ansatzweise möglich ist. Bei den Merkmalen *Jobben, Miete* und *Geschwister* sinkt die Korrelation. Bei den übrigen Merkmalen ist zwar eine Verbesserung der Ergebnisse feststellbar, allerdings fällt die Erhöhung der Korrelation sehr gering aus. Eine Ausnahme bildet das Merkmal *Altersunterschied*, bei dem ein deutlicher Zuwachs zu beobachten ist. Allerdings

wurde gerade dieses Merkmal zur Transformation der Ratingskala herangezogen. Insofern nimmt es bei der Interpretation eine Sonderrolle ein.

| | $r_{x_j y_j}^{\tau\text{-b}}$ | $r_{x_j^{neu} y_j}^{\tau\text{-b}}$ |
|---|---|---|
| Vorlesung | 0,420 | 0,430 |
| Lernstunden | 0,160 | 0,163 |
| Jobben | 0,554 | *0,544* |
| Einkommen | 0,255 | 0,256 |
| Geld | 0,300 | 0,306 |
| Uni | 0,365 | 0,367 |
| Ehrenamt | 0,521 | 0,524 |
| Entfernung | 0,328 | 0,335 |
| Hobbys | 0,385 | 0,388 |
| Miete | 0,528 | *0,525* |
| Wohnfläche | 0,448 | 0,461 |
| Geschwister | 0,760 | *0,719* |
| Altersunterschied | 0,584 | 0,627 |
| Schlaf | 0,403 | 0,408 |

**Tabelle 6.19: Korrelation bei Transformation auf Basis des Merkmals Altersunterschied**

Es kann somit durch eine Erhebung lediglich eines zusätzlichen Merkmals keine deutliche Verbesserung erreicht werden. Die gezeigten Verbesserungen sind nur minimal bzw. teilweise verschlechtern die Ergebnisse sich sogar. Die Ergebnisse liefern somit keine Rechtfertigung für den Anstieg im Erhebungsaufwand, den diese Methode nach sich ziehen würde.

## 6.3 Kritische Würdigung der Ergebnisse

Die Ergebnisse der empirischen Studie belegen, dass 5 Antwortkategorien vor dem Hintergrund der Kriteriumsvalidität optimal sind. Darüber hinaus legen die Resultate den Schluss nahe, dass inhomogene Mappingstrategien für eine

Verschlechterung der Qualität der Datenerhebung mittels Ratingskalen verantwortlich sind. Außerdem kann die Identifikation dieser Abbildungsstrategien zu einer Verbesserung der Messung beitragen.

Die eigene Studie erweitert somit den Kenntnisstand zur Ausgestaltung einer Ratingskala. Dennoch bestehen immer noch Aspekte bei der Konstruierung einer Ratingskala, die im Rahmen dieser Arbeit nicht beleuchtet wurden, die es unter Umständen jedoch zukünftig zu erforschen gilt.

- Die vorgestellte eigene Studie ermöglicht keine Aussagen über die Verwendung der Kategorie ´Keine Meinung´. Die existierenden Studien zu diesem Thema überprüfen in erster Linie die Auswirkungen auf die Reliabilität. Wie die eigene empirische Studie jedoch gezeigt hat, kann die Berücksichtigung der Validität jedoch zu anderen Ergebnissen führen.

- Die eigene empirische Studie konzentriert sich auf ungerade Anzahl Antwortkategorien und entspricht damit einer langen Tradition von Autoren, die eine ungerade Anzahl bevorzugen (vgl. KROSNICK, FABRIGAR, 1997, S. 146). Dennoch erscheint ein Vergleich ungerader Antwortkategorien mit geradzahligen Antwortmöglichkeiten sinnvoll.

Das größte offene Forschungsfeld ist aber zweifelsfrei die Integration der psychologischen Prozesse in die Überlegungen, welche Anzahl an Antwortkategorien optimal ist. Wenn es beispielsweise gelingt, idealtypische Abbildungsprozesse zu identifizieren und zu steuern, kann die Qualität der Messergebnisse deutlich verbessert werden. Dazu ist jedoch vor allem die Erforschung der verschiedenen psychologischen Prozesse erforderlich. Die bisherigen Arbeiten zu diesem Thema versuchen zwar verschiedene Befragungsphänomene modellartig zu erklären, bis auf den Beantwortungsprozess von TOURANGEAU et al. (2005) existiert allerdings kein ganzheitliches Modell, das tatsächlich alle Aspekte psychologischer Prozesse zu integrieren versucht.

# Kapitel 7

# Fazit

In dieser Arbeit stand die Messung im wirtschafts- und sozialwissenschaftlichen Bereich im Vordergrund. Befragungen ermöglichen es Unternehmen, Wissen über Konsumenten zu erlangen. Dabei tritt für den Messenden jedoch häufig das Problem auf, dass interessierende Eigenschaften, Einstellungen und Überzeugungen nicht direkt erfassbar sind. Ratingskalen stellen die bedeutsamste Fragetechnik zur Erhebung solcher Daten dar.

Im Rahmen dieser Arbeit sollte gezeigt werden, wie viele Antwortkategorien einer Ratingskala optimal sind. Darüber hinaus sollte beurteilt werden, ob die Messergebnisse einer Ratingskala durch die Identifikation des Abbildungsprozesses verbessert werden können.

Die bestehenden Arbeiten zur Gestaltung einer Ratingskala stellen vor allem deren *Reliabilität* in den Mittelpunkt. Die neueren Studien zeigen, dass unter dem Gesichtspunkt der Reliabilität 7 oder mehr Antwortkategorien ideal sind.

Die eigene empirische Studie überprüft, ob dieses Ergebnis auch bei der Berücksichtigung der *Validität* von Ergebnissen zu beobachten ist. Wie die Ausführungen in Kapitel 6.2.1 gezeigt haben, bestätigen sich die Ergebnisse von PRESTON und COLMAN (2000) nicht. Die 7-stufige Ratingskala schneidet im Vergleich zu den 3- und 5-stufigen Ratingskalen deutlich schlechter ab.

Insbesondere inhomogene Mappingprozesse sind dafür verantwortlich, dass Ratingskalen mit weniger Antwortkategorien unter dem Gesichtspunkt der Validität zu bevorzugen sind. Die 5-stufige Ratingskala liefert dabei etwas bessere Ergebnisse, als die 3-stufige Ratingskala.

Der Versuch die Ergebnisse durch eine Identifikation der Mappingprozesse zu verbessern führt zu keinen in der Praxis verwertbaren Ergebnissen. Für die

Existenz dieser inhomogenen Mappingprozesse liefert Kapitel 6.2.2 zwar mehrere Hinweise und prinzipiell scheint diese Überlegung auch viel versprechend zu sein, wie die Verwendung aller Ratingskalen und externer Kriterien in Kapitel 6.2.3.1 gezeigt hat. Allerdings mangelt es diesem Ansatz an der Umsetzbarkeit in der Praxis. Die für die Praxis eher geeigneten Methoden, die auf ein externes Kriterium völlig verzichten (Kapitel 6.2.3.2) oder nur ein externes Kriterium erheben (Kapitel 6.2.3.3) sorgen wiederum nicht für die erhoffte Verbesserung der Ratingskalen.

Unter Berücksichtigung der eigenen Studie und der bisherigen Forschungsergebnisse können folgende Aussagen zur Anzahl der Antwortkategorien abschließend festgehalten werden:

- Die Reliabilität der Ratingskala nimmt zu, wenn 7 oder mehr Antwortkategorien verwendet werden.
- Die Validität der Ratingskala ist bei 5 Antwortkategorien am höchsten.

Hinsichtlich der konkreten Ausgestaltung einer Ratingskala sollten zukünftig folgende Fragestellungen unterschieden werden:

1. Soll eine Person wiederholt eine Frage zum selben Thema beantworten?
2. Sollen mehrere Personen die gleiche Frage beantworten?

Die erste Fragestellung ist typisch für Panel-Untersuchungen, für Fragen zur Einstellungsänderung von Personen oder zur Werbewirksamkeit von Marketingmaßnahmen. Das Ziel ist dabei herauszufinden, wie sich die Ergebnisse für eine Person im Zeitablauf verändern. Für diese Art der Fragestellungen ist insbesondere die zeitliche Konsistenz und somit die Reliabilität der Ergebnisse entscheidend. Deshalb sollten 7 oder mehr Antwortkategorien verwendet werden.

Bei der zweiten Fragestellung steht ein Vergleich der Antworten mehrerer Personen im Fokus. Dies geschieht immer dann, wenn es Ziel der Studie ist, zu erfahren, welche Personen von einem Produkt überzeugt sind, ein bestimmtes Flugticket kaufen oder nicht, einen Kredit zurückbezahlen können und vieles mehr. Die Vergleichbarkeit einzelner Probanden setzt hierbei voraus, dass die Antworten keiner systematischen Verzerrung, ausgelöst z.B. durch inhomogene

Mappingprozesse, unterliegen. Für diese Art der Fragestellung sollten deshalb 5 oder noch weniger Antwortalternativen gewählt werden.

Die optimale Anzahl der Antwortkategorien kann also nicht allgemeingültig für alle Fragestellungen festgelegt werden. Zur Vermeidung von Fehlinterpretationen, sollte die Ratingskala vor dem individuellen Befragungshintergrund ausgestaltet werden. Nur so lassen sich möglichst sinnvolle Informationen als Basis für unternehmerische Strategien und Entscheidungen generieren.

# Literaturverzeichnis


- **Aaker, D.A.; Day, G.S. (1990)**: Marketing Research; Wiley, New York
- **Acock, A.C.; Martin D.J. (1974)**: The Undermeasurement Controversy: Should Ordinal Data be Treated as Interval?; in: Sociology and Social Research, 58 (4), S. 427-433
- **Allen, M.J.; Yen, W.M. (1979)**: Introduction to Measurement Theory; Brooks, Monterey
- **Allport, G.W. (1935)**: Attitudes; in: Murchison, C. (Hrsg): A handbook of social psychology, S. 798-844; Clark University Press, Worcester
- **Alwin, D.F.; Krosnick, J.A. (1991)**: The Reliability of Survey Attitude Measurement: The Influence of Question and Respondent Attributes; in: Sociological Methods and Research, 20, S. 139-181
- **Andrews, F.M. (1984)**: Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach; in: The Public Opinion Quarterly, 48 (2) (Summer, 1984), S. 409-442
- **Backhaus, K.; Erichson, B.; Plinke, W.; Weiber, R. (2006)**: Multivariate Analysemethoden; Springer, Berlin
- **Bamberg, G.; Baur, F. (2002)**: Statistik; Oldenbourg, München
- **Belson, W.A. (1981)**: The design and understanding of survey questions; Gower, Aldershot
- **Berekoven, L.; Eckert, W.; Ellenrieder, P. (1989)**: Marktforschung; Gabler, Wiesbaden
- **Bickart, B.A.; Blair, J.; Menon G.; Sudman, S. (1990)**: Cognitive aspects of proxy reporting of behaviour; in: Advances in Consumer Research, 17, S. 198-206
- **Bishop, G.; Oldenburg, R.; Tuchfarber, A. (1986)**: Opinions on fictitious issues: The pressure to answer survey questions; in: Public Opinion Quarterly, 50, S. 240-250
- **Blair, E.A.; Burton, S. (1987)**: Cognitive processes used by survey respondents to answer behavirol frequency questions; in: Journal of Consumer Research, 14, S. 280-288
- **Bosch, K. (1998)**: Statistik-Taschenbuch, Oldenbourg, München
- **Bradburn, N.M.; Sudman, S. (1979)**: Improving interview method and questionnaire design; Jossey-Bass, San Francisco
- **Burton, S.; Blair, E.A. (1991)**: Task conditions, response formulation processes, and response accuracy for behavirol frequency questions in surveys; in: Public Opinion Quarterly, 55, S. 50-79
- **Cannell, C.; Miller, P.; Oksenberg, L. (1981)**: Research on interviewing techniques; in: Leinhardt, S. (Hrsg): Sociological methodology, S. 389-437, Jossey Bass, San Francisco

- **Collins, A.M.; Michalski, R. (1989)**: The logic of plausible reasoning: A core theory; in: Cognitive Science, 13, S. 1-50

- **Conrad, F.G.; Brown, N.R.; Cashman, E.R. (1998)**: Strategies for estimating behavioral frequency in survey interviews; in: Memory, 6, S. 339-366

- **Conway, M.A. (1996)**: Autobiographical Knowledge and autobiographical memories; in: Rubin, D.C. (Hrsg): Remembering our past, S. 67-93, Cambridge University Press, Cambridge

- **Coombs, C.H. (1983)**: Psychology and Mathematics; Wiley, Rexdale

- **Cox, E.P. (1980)**: The optimal number of response alternatives for a scale: A review; in: Journal of Marketing Research, 17, S. 407-422

- **Daamen, D.D.L.; de Bie, S.E. (1992)**: Serial context effects in survey items; in: Schwarz, N.; Sudman, S. (Hrsg): Context effects in social and psychological research, S. 97-114, Springer, New York

- **Denz, H. (1976)**: Trennschärfebestimmung von Items und Likert-Skalierung; in: Holm, K. (Hrsg.): Die Befragung 4, S. 96-105, Francke, München

- **Eid, M.; Diener, E. (2006)**: Introduction: The need for multimethod measurement in psychology; in: Eid, M.; Diener, E. (Hrsg.): Handbook of psychological measurement: A multimethod perspective, S. 3-8, American Psychological Association, Washington DC

- **Eid, M.; Lischetzke, T.; Nussbeck, F. W. (2006)**: Structural equation models for multitrait-multimethod data; in: Eid, M.; Diener, E. (Hrsg.): Handbook of psychological measurement: A multimethod perspective, S. 283-299, American Psychological Association, Washington DC

- **Fazio, R. (1989)**: On the power and functionality of attitudes: The role of attitude accessibility; in: Pratkanis, A.; Breckler, S.; Greenwald, A. (Hrsg): Attitude structure and function, S. 153-179, Erlbaum, Hillsdale

- **Finn, R.H. (1972)**: Effects of some Variations in Rating Scale Characteristics on the Means and Reliabilities of Ratings; in: Journal of Educational and Psychological Measurement, 32, S. 255-265

- **Fishbein, M. (1963)**: An Investigation of the Relationships between Beliefs about an Object and the Attitude Toward that Object; in: Human Relations, 16, S. 233-239

- **Gierl, H. (1995)**: Marketing; Kohlhammer, Stuttgart

- **Gillund, G.; Shiffrun, R.M. (1984)**: A retrieval model for both recognition and recall; in: Psychological Review, 91, S. 1-67

- **Graesser, A.C.; Bommareddy, S.; Swamer, S.; Golding, J.M. (1996)**: Integrating questionnaire design with a cognitive computational model of human question answering; in: Schwarz, N.; Sudman, S. (Hrsg): Answering questions, S. 143-174, Jossey-Bass, San Francisco

- **Hair, J.F.; Black, B.; Babin, B.; Anderson, R.E.; Tatham, R.L. (2006)**: Multivariate Data Analysis; Pearson, Upper Saddle River

- **Hamerle, A. (1982)**: Latent-Trait-Modelle; Beltz, Weinheim
- **Hammann, P.; Erichson B. (2000)**: Marktforschung; Fischer, Stuttgart
- **Handl, A. (2002)**: Multivariate Analysemethoden; Springer, Berlin
- **Hartung, J.; Elpelt, B. (1992)**: Multivariate Statistik; Oldenbourg, München
- **Higgins, T.E.; McCann, D.C. (1984)**: Social Encoding and Subsequent Attitudes, Impressions, and Memory: "Context-Driven" and Motivational Aspects of Processing; in: Journal of Personality and Social Psychology, 47 (1), S. 26-39
- **Hilbert, A. (1998)**: Zur Theorie der Korrelationsmaße; Eul, Lohmar
- **Hippler, H.M.; Schwarz, N. (1987)**: Response Effects in Survey; in: Hippler, H.M.; Schwarz, N.; Sudman, S. (Hrsg.): Social Information Processing and Survey Methodology, S. 84-100, Springer, New York
- **Holm, K. (1975)**: Die Frage; in: Holm, K. (Hrsg.): Die Befragung 1, S. 32-91, Francke, München
- **Huttenlocher, J.; Hedges, L.V.; Bradburn, N.M. (1990)**: Reports of elapsed time: Bounding and rounding processes in estimation; in: Journal of Experimental Psychology: Learning, Memory, and Cognition, 16, S. 196-213
- **Janssen, J.; Laatz, W. (2005)**: Statistische Datenanalyse mit SPSS für Windows; Springer, Berlin
- **Judd, C.; Krosnick, J. (1982)**: Attitude centrality, organization, and measurement; in: Journal of Personality and Social Psychology, 42, S. 436-447
- **Judd, C.; Milburn, J. (1980)**: The structure of attitude systems in the general public: Comparision of a structural equation model; in: American Sociological Review, 46, S. 660-669
- **Kallmann, A. (1979)**: Skalierung in der empirischen Forschung; V. Florentz, München
- **Kendall, M.G. (1970)**: Rank Correlation Methods; Griffin, London
- **Klein, R; Scholl A. (2004)**: Planung und Entscheidung; Vahlen, München
- **Kolodner, J. (1985)**: Memory of experience; in: Bower, G.H. (Hrsg): The psychology of learning and motivation; S. 1-57, Academic Press, Orlando
- **Krantz, D.H.; Luce, R.D.; Suppes, P.; Tversky, A. (1971)**: Foundations of Measurement, Bd. 1; Akademic Press, New York
- **Kreppner, K. (1975)**: Zur Problematik des Messens in den Sozialwissenschaften; Klett, Stuttgart
- **Kroeber-Riel, W.; Weinberg, P. (1999)**: Konsumentenverhalten; Vahlen, München
- **Krosnick, J.A. (1991)**: Response Strategies for Coping with the Cognitive Demands of Attitude Measures in Surveys; in: Applied Cognitive Psychology, 5, S. 213-236

- **Krosnick, J.A.; Alwin, D. (1987)**: An evaluation of a cognitive theory of response-order effects in survey measurement; in: Public Opinion Quarterly, 51, 201-219

- **Krosnick, J. A.; Berent, M. K. (1993)**: Comparisons of party identification and policy preferences: The impact of survey question format; in: American Journal of Political Science, 37, S. 941-964

- **Krosnick,J.A.; Fabrigar, L. R. (1997)**: Designing rating scales for effective measurement in surveys; in: Lyberg, L.; Biemer, P.; Collins, M.; Decker, L.; DeLeeuw, E.; Dippo, C.; Schwarz N.; Trewin, D. (Hrsg.): Survey Measurement and Process Quality, S. 141-150, Wiley, New York

- **Krosnick, J. A.; Schuman, H. (1988)**: Attitude intensity, importance, and certainty and susceptibility to response effects; in: Journal of Personality and Social Psychology, 54, S. 940-952

- **Kruskal, J.B. (1964)**: Multidimensional Scaling by Optimizing Goodness of Fit to a Non-metric Hypothesis; in: Psychometrika, 29, S. 1-28

- **Labovitz, S. (1967)**: Some Observations on Measurement and Statistics; in: Social Forces, 46, S. 151-160

- **Landy, F.J.; Farr, J.L. (1980)**: Performance rating; in: Psychological Bulletin, 87, S. 72-107

- **Lessler, J.T.; Tourangeau, P.; Salter, W. (1989)**: Questionnaire design in the cognitive research laboratory: Results of an experimental prototype; in: Vital and Health Statistics, Band 6, Nr. 1, U.S. Government Printing Office, Washington

- **Lodge, M.; McGraw, K.; Stroh, P. (1989)**: An impression-driven model of candidate evaluation; in: American Political Science Review, 83, S. 399-419

- **Luce, R.D.; Krantz, D.H.; Suppes, P.; Tversky, A. (1990)**: Foundations of Measurement, Bd. 3; Akademic Press, San Diego

- **Mayntz, R.; Holm, K.; Hübner, P. (1972)**: Einführung in die Methoden der empirischen Soziologie; Westd. Verlag, Opladen

- **Meffert, H. (1992):** Marketingforschung und Käuferverhalten; Gabler, Wiesbaden

- **McClendon, M.; O´Brien, D. (1988)**: Question-order effects on subjective well-being; in: Public Opinion Quarterly, 52, S. 351-364

- **Moxey, L.M.; Sanford, A.J. (1993)**: Communicating quantities; Erlbaum, Hillsdale

- **Narayan, S.; Krosnick, J. A. (1996)**: Education moderates some response effects in attitude measurement; in: Public Opinion Quarterly, 60, S. 58-88

- **Nelson, E.A.; Grinder, R.E.; Mutterer, M.L. (1969)**: Sources of Variance in Behaviour Measures of Honesty in Temptation Situations; in: Developmental Psychology, 21, S. 265-279

- **Neter, J.; Waksberg, J. (1964)**: A study of response errors in expenditures data from household interviews; in: Journal of American Statistical Association, 59, S. 17-55

- **Opitz, O. (1980)**: Numerische Taxonomie; Gustav Fischer, Stuttgart
- **Opitz, O. (2004)**: Mathematik; Oldenbourg, München
- **Osgood, C.E.; Suci, G.J.; Tannenbaum, P.H. (1957)**: The Measurement of Meaning, Urbana, Chicago
- **Orth, B. (1974)**: Einführung in die Theorie des Messens; Kohlhammer, Stuttgart
- **Ottati, V.; Riggle, E.; Wyer, R.; Schwarz, N.; Kuklinski, J. (1989)**: Cognitive and affective bases of opinion survey responses; in: Journal of Personality and Social Psychology, 57, S. 404-415
- **Parducci, A. (1965)**: Category Judgment: A range-frequency model; in: Psychological Review, 72, S. 407-418
- **Parducci, A. (1974)**: Contextual effets: A range-frequency analysis; in: Carterette, E.; Friedman, M. (Hrsg): Handbook of perception: Psychophysical judgment and measurement, S. 127-141, Academic Press, New York
- **Pfanzagl, J. (1968)**: Theory of Measurement; Physica, Würzburg
- **Poulton, E.C. (1989)**: Bias in quantifying judgments; Erlbaum, Hillsdale
- **Preston, C.C.; Colman, A.M. (2000)**: Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences; in Acta Psychologica, 104, S. 1-15
- **Rogge, H.J. (1981)**: Marktforschung; Hanser, München
- **Rosenberg, M.J. (1967)**: Cognitive Structure and Attitudinal Affect; in: Fishbein, M. (Hrsg.): Readings in Attitude Theory and Measurement, S. 325-342, Wiley, New York
- **Sanbonmatsu, D.; Fazio, R. (1990)**: The role of attitudes in memory-based decision-making; in: Journal of Personality and Social Psychology, 59, S. 614-622
- **Schmitt, M. (2006)**: Conceptual, Theoretical and Historical Foundations of Multimethod Assessment; in: Eid, M.; Diener, E. (Hrsg.): Handbook of psychological measurement: A multimethod perspective, S. 9-25, American Psychological Association, Washington DC
- **Schnell, R.; Hill, P.B.; Esser, E. (2005)**: Methoden der empirischen Sozialforschung; Oldenbourg, München
- **Schuman, H., Presser, S. (1981)**: Questions and answers in attitude surveys: Experiments in question form, wording, and context; Academic Press, New York
- **Schwarz, N.; Hippler, H.J. (1987)**: What response scales may tell your respondents: Information functions of response alternatives; in: Hippler, H.J.; Schwarz, N.; Sudman, S. (Hrsg): Social information processing and survey methodology, S. 163-178, Springer, New York

- **Schwarz, N.; Hippler, H.J.; Deutsch, B.; Strack, F. (1985)**: Response Scales: Effects of category range on reported behaviour and comparative judgments; in: Public Opinion Quarterly, 49, S. 388-395

- **Schwarz, N.; Hippler, H.; Noelle-Neumann, E. (1991)**: A cognitive model of response-order effects in survey measurement; in: Schwarz, N.; Sudman, S. (Hrsg): Context effects in social and psychological research, S. 187-201, Springer, New York

- **Searle, J. (1969)**: Speech acts; Cambridge University Press, Cambridge

- **Sears, D.O. (1983)**: The person-positivity bias; in: Journal of Personality and Social Psychology, 44, S. 233-250

- **Shepard, R.N. (1962)**: The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function I; in: Psychometrika, 27, S. 125-140

- **Shryock, H.S.; Siegel, J.S.; Stockwell, E.G. (1976)**: The methods and materials of demography; Academic Press, San Diego

- **Simon, H. (1957)**: Models of man; Wiley, New York

- **Sixtl, F. (1976)**: Skalierungsverfahren: Grundzüge und ausgewählte Methoden sozialwissenschaftlichen Messens; in: Holm, K. (Hrsg.): Die Befragung 4, S. 1-86, Francke, München

- **Smith, A.F. (1991)**: Cognitive Processes in long-term dietary recall; in: Vital and Health Statistics, Band 6, Nr. 4, U.S. Government Printing Office, Washington

- **Smith, E.R. (1999)**: New connectionist models of mental representation: Implications for survey research; in: Sirken, M.G.; Herrmann, D.J.; Schechter, S.; Schwarz, N.; Tanur, J.M.; Tourangeau, R. (Hrsg): Cogniton and survey research, S. 251-266, Wiley, New York

- **Stevens, S.S. (1946)**: On The Theory of Scales of Measurement; in: Science, 103, S. 677-680

- **Stevens, S.S. (1951)**: Mathematics, Measurement and Psychophysics; in: Stevens, S.S. (Hrsg.): Handbook of Experimental Psychology, Wiley, New York

- **Strack, F.; Martin, L. (1987)**: Thinking, judging, and communicating: A process account of context effects in attitude surveys; in: Hippler, H.; Schwarz, N.; Sudman, S. (Hrsg): Social information processing and survey methodology, S. 123-148, Springer, New York

- **Sudman, S.; Bradburn, N. (1973)**: Effects of time and memory factors on response in surveys; in: Journal of the American Statistical Association, 68, S. 805-815

- **Thurstone, L. (1927)**: A law of comparative judgment; in: Psychological Review, 34, S. 273-286

- **Torgerson, W.S. (1958)**: Theory and Methods of Scaling; Wiley, New York

- **Tourangeau R.; Rasinski, K. (1988)**: Cognitive Processes underlying context effects in attitude measurement; in: Psychological Bulletin, 103, S. 299-314

- **Tourangeau, R.; Rips, L.J.; Rasinski, K. (2005)**: The psychology of survey response; Cambridge University Press, Cambridge

- **Tourangeau, R.; Smith, T.W.; Rasinski, K.A. (1997)**: Motivation to report sensitive behaviors on surveys: Evidence from a bogus pipeline experiment; in: Journal of Applied Social Psychology, 27, S. 209-222

- **Trochim, W.M. (2007)**: The Research Methods Knowledge Base, in: http://www.socialresarchmethods.net/kb/relandval.php, Zugriff am 21.12.2007

- **Trommsdorff, V. (1989)**: Konsumentenverhalten; Kohlhammer, Stuttgart

- **Tulving, E. (1983)**: Elements of episodic memory; Oxford University Press, Oxford

- **Wallsten, T.S.; Budescu, T.V.; Zwick, R.; Kemp, S.M. (1993)**: Preferences and reasons for communicating probabilistic information in numerical or verbal terms; in: Bulletin of the Psychonomic Society, 31, S. 135-138

- **Wedell, D.H. (1990)**: Method for determining the focus of context effects in judgements; in Caverni, J.P.; Fabre, J.M.; Gonzalez, M. (Hrsg.): Cognitive Biases, S. 285-304, Elsevier Science, Amsterdam

- **Wettschurek, G. (1974)**: Indikatoren und Skalen in der demoskopischen Marktforschung; in: Behrens, K.C. (Hrsg.): Handbuch der Marktforschung, S. 285-324, Gabler, Wiesbaden

- **Willis, G.; Brittingham, A.; Lee, L.; Tourangeau, R.; Ching, P. (1999)**: Response errors in surveys of children´s immnizations; in: Vital and Health Statistics, Band 6, Nummer 8, National Center for Health Statistics, Hyattsville

- **Wilson, T.D.; Hodges, S. (1992)**: Attitudes as temporary constructions; in: Martin, L.; Tesser, A. (Hrsg): The construction of social judgments, S. 37-66, Springer, New York

- **Zajonc, R.B. (1968)**: Cognitive Theories in social psychology; in: Lindzey, G.; Aronson, E. (Hrsg): Handbook of social psychology, S. 320-411, Addison-Wesley, Reading

- **Zaller, J.R. (1992)**: The nature and origins of mass opinion; Cambridge University Press, Cambridge

# Anhang

Vielen Dank für die Bereitschaft zur Teilnahme an dieser Befragung. Dieser Fragebogen dient ausschließlich einem wissenschaftlichen Zweck. Alle erhobenen Daten werden selbstverständlich anonym und vertraulich behandelt.

| | | |
|---|---|---|
| Welches Geschlecht haben Sie? | Männlich ☐ | Weiblich ☐ |
| Haben Sie Ihr Vordiplom bereits bestanden? | Ja ☐ | Nein ☐ |
| Wie viele Geschwister haben Sie? | | |
| In welchem Jahr wurde Ihre Mutter geboren? | | |
| In welchem Jahr wurde Ihr Vater geboren? | | |
| Wie viele Stunden Schlaf benötigen Sie, um sich fit zu fühlen? | | |
| Wie groß sind Sie (in cm)? | | |
| Wie hoch ist ungefähr Ihr monatliches Einkommen (in Euro) **vor** Abzug fixer Kosten (Lohn, Taschengeld, BAföG, …)? | | |
| Wie viel Geld (in Euro) steht Ihnen monatlich zur freien Verfügung? | | |
| Wie viel Miete bezahlen Sie monatlich (Euro)? | | |
| Wie weit entfernt von der Universität wohnen Sie (km)? | | |
| Wie viele Quadratmeter Wohnfläche stehen ausschließlich Ihnen zur Verfügung? | | |
| Wie viele Stunden jobben Sie regelmäßig während des Semesters pro Woche? | | |
| Wie viel Zeit (in Stunden) nehmen Sie sich wöchentlich für Hobbys (Sport, Musik, …)? | | |
| Wie viele Stunden arbeiten Sie pro Woche ehrenamtlich (Rotes Kreuz, Sportverein, …)? | | |
| An wie vielen Tagen (pro Semester) engagieren Sie sich außerhalb des Studiums im engeren Sinn für die Universität (StuRa, AStA, ...)? | | |
| Wie viele Stunden bereiten Sie, abgesehen von der direkten Prüfungsvorbereitung, die Vorlesungen wöchentlich vor bzw. nach? | | |
| Wie viele Stunden lernen Sie in der Vorbereitungszeit auf die Prüfungen durchschnittlich am Tag? | | |

Wenn Sie sich mit **anderen Studenten vergleichen**, wie sehr würden Sie dann folgenden Aussagen zustimmen? Beziehen Sie sich dabei bitte auf Ihre **derzeitige Situation**!

| | Stimme zu | Teils teils | Stimme nicht zu |
|---|---|---|---|
| Ich verwende während des Semesters viel Zeit auf die Vor- und Nachbereitung von Vorlesungen | □ | □ | □ |
| Mein tägliches Lernpensum ist in der Vorbereitungszeit auf die Prüfungen größer als bei anderen Studenten | □ | □ | □ |
| Während des Semesters jobbe ich überdurchschnittlich viel | □ | □ | □ |
| Für einen Studenten ist mein monatliches Einkommen (Lohn, Taschengeld, ...) relativ hoch (vor Abzug der Fixkosten) | □ | □ | □ |
| Für einen Studenten habe ich sehr viel Geld zur freien Verfügung | □ | □ | □ |
| Mein ehrenamtliches Engagement für die Universität (StuRa, AStA, ...) ist groß | □ | □ | □ |
| Mein ehrenamtliches Engagement außerhalb der Universität (Rotes Kreuz, Sportverein, …) ist groß | □ | □ | □ |
| Ich habe eine weite Anfahrt von meinem Wohnort zur Universität | □ | □ | □ |
| Ich verwende überdurchschnittlich viel Zeit für Hobbys (Sport, Musik, …) | □ | □ | □ |
| Ich bezahle eine vergleichsweise hohe Miete | □ | □ | □ |
| Ich habe sehr viel Wohnraum für mich allein | □ | □ | □ |
| Im Vergleich zu anderen habe ich viele Geschwister | □ | □ | □ |
| Der Altersunterschied meiner Eltern ist groß | □ | □ | □ |
| Ich benötige viel Schlaf, um mich fit zu fühlen | □ | □ | □ |
| Ich bin überdurchschnittlich groß | □ | □ | □ |

Beantworten Sie jetzt bitte die Fragen nochmals: Geben Sie jetzt **Ihre Einschätzung** ab, welche Werte für einen **durchschnittlichen Studenten** gelten:

| | |
|---|---|
| Wie hoch ist ungefähr das monatliche Einkommen in Euro **vor** Abzug fixer Kosten (Lohn, Taschengeld, BAföG, …)? | |
| Wie viel Geld (in Euro) steht monatlich zur freien Verfügung? | |
| Wie viel Miete wird monatlich bezahlt (Euro)? | |
| Wie weit ist die Entfernung von der Universität zum Wohnort (km)? | |
| Wie viele Quadratmeter Wohnfläche stehen ausschließlich zur eigenen Verfügung? | |
| Wie viele Stunden wird während des Semesters pro Woche ungefähr gejobbt? | |
| Wie viel Zeit (in Stunden) wird wöchentlich für Hobbys (Sport, Musik, …) verwendet? | |
| Wie viele Stunden pro Woche wird ehrenamtlich gearbeitet (Rotes Kreuz, Sportverein, …)? | |
| An wie vielen Tagen (pro Semester) erfolgt ein Engagement für die Universität außerhalb des Studiums im engeren Sinn (StuRa, AsTa, etc.)? | |
| Wie viele Stunden bereitet man, abgesehen von der direkten Prüfungsvorbereitung, die Vorlesungen wöchentlich vor bzw. nach? | |
| Wie viele Stunden lernt man in der Vorbereitungszeit auf die Prüfungen durchschnittlich am Tag? | |

Vielen Dank für die Teilnahme an dieser Umfrage!

Vielen Dank für die Bereitschaft zur Teilnahme an dieser Befragung. Dieser Fragebogen dient ausschließlich einem wissenschaftlichen Zweck. Alle erhobenen Daten werden selbstverständlich anonym und vertraulich behandelt.

| | | |
|---|---|---|
| Welches Geschlecht haben Sie? | Männlich ☐ | Weiblich ☐ |
| Haben Sie Ihr Vordiplom bereits bestanden? | Ja ☐ | Nein ☐ |
| Wie viele Geschwister haben Sie? | | |
| In welchem Jahr wurde Ihre Mutter geboren? | | |
| In welchem Jahr wurde Ihr Vater geboren? | | |
| Wie viele Stunden Schlaf benötigen Sie, um sich fit zu fühlen? | | |
| Wie groß sind Sie (in cm)? | | |
| Wie hoch ist ungefähr Ihr monatliches Einkommen (in Euro) **vor** Abzug fixer Kosten (Lohn, Taschengeld, BAföG, …)? | | |
| Wie viel Geld (in Euro) steht Ihnen monatlich zur freien Verfügung? | | |
| Wie viel Miete bezahlen Sie monatlich (Euro)? | | |
| Wie weit entfernt von der Universität wohnen Sie (km)? | | |
| Wie viele Quadratmeter Wohnfläche stehen ausschließlich Ihnen zur Verfügung? | | |
| Wie viele Stunden jobben Sie regelmäßig während des Semesters pro Woche? | | |
| Wie viel Zeit (in Stunden) nehmen Sie sich wöchentlich für Hobbys (Sport, Musik, …)? | | |
| Wie viele Stunden arbeiten Sie pro Woche ehrenamtlich (Rotes Kreuz, Sportverein, …)? | | |
| An wie vielen Tagen (pro Semester) engagieren Sie sich außerhalb des Studiums im engeren Sinn für die Universität (StuRa, AStA, ...)? | | |
| Wie viele Stunden bereiten Sie, abgesehen von der direkten Prüfungsvorbereitung, die Vorlesungen wöchentlich vor bzw. nach? | | |
| Wie viele Stunden lernen Sie in der Vorbereitungszeit auf die Prüfungen durchschnittlich am Tag? | | |

Wenn Sie sich mit **anderen Studenten vergleichen**, wie sehr würden Sie dann folgenden Aussagen zustimmen? Beziehen Sie sich dabei bitte auf Ihre **derzeitige Situation**!

| | Stimme zu | Stimme eher zu | Teils teils | Stimme eher nicht zu | Stimme nicht zu |
|---|---|---|---|---|---|
| Ich verwende während des Semesters viel Zeit auf die Vor- und Nachbereitung von Vorlesungen | □ | □ | □ | □ | □ |
| Mein tägliches Lernpensum ist in der Vorbereitungszeit auf die Prüfungen größer als bei anderen Studenten | □ | □ | □ | □ | □ |
| Während des Semesters jobbe ich überdurchschnittlich viel | □ | □ | □ | □ | □ |
| Für einen Studenten ist mein monatliches Einkommen (Lohn, Taschengeld, ...) relativ hoch (vor Abzug der Fixkosten) | □ | □ | □ | □ | □ |
| Für einen Studenten habe ich sehr viel Geld zur freien Verfügung | □ | □ | □ | □ | □ |
| Mein ehrenamtliches Engagement für die Universität (StuRa, AStA, ...) ist groß | □ | □ | □ | □ | □ |
| Mein ehrenamtliches Engagement außerhalb der Universität (Rotes Kreuz, Sportverein, …) ist groß | □ | □ | □ | □ | □ |
| Ich habe eine weite Anfahrt von meinem Wohnort zur Universität | □ | □ | □ | □ | □ |
| Ich verwende überdurchschnittlich viel Zeit für Hobbys (Sport, Musik, …) | □ | □ | □ | □ | □ |
| Ich bezahle eine vergleichsweise hohe Miete | □ | □ | □ | □ | □ |
| Ich habe sehr viel Wohnraum für mich allein | □ | □ | □ | □ | □ |
| Im Vergleich zu anderen habe ich viele Geschwister | □ | □ | □ | □ | □ |
| Der Altersunterschied meiner Eltern ist groß | □ | □ | □ | □ | □ |
| Ich benötige viel Schlaf, um mich fit zu fühlen | □ | □ | □ | □ | □ |
| Ich bin überdurchschnittlich groß | □ | □ | □ | □ | □ |

Beantworten Sie jetzt bitte die Fragen nochmals: Geben Sie jetzt **Ihre Einschätzung** ab, welche Werte für einen **durchschnittlichen Studenten** gelten:

| | |
|---|---|
| Wie hoch ist ungefähr das monatliche Einkommen in Euro **vor** Abzug fixer Kosten (Lohn, Taschengeld, BAföG, …)? | |
| Wie viel Geld (in Euro) steht monatlich zur freien Verfügung? | |
| Wie viel Miete wird monatlich bezahlt (Euro)? | |
| Wie weit ist die Entfernung von der Universität zum Wohnort (km)? | |
| Wie viele Quadratmeter Wohnfläche stehen ausschließlich zur eigenen Verfügung? | |
| Wie viele Stunden wird während des Semesters pro Woche ungefähr gejobbt? | |
| Wie viel Zeit (in Stunden) wird wöchentlich für Hobbys (Sport, Musik, …) verwendet? | |
| Wie viele Stunden pro Woche wird ehrenamtlich gearbeitet (Rotes Kreuz, Sportverein, …)? | |
| An wie vielen Tagen (pro Semester) erfolgt ein Engagement für die Universität außerhalb des Studiums im engeren Sinn (StuRa, AsTa, etc.)? | |
| Wie viele Stunden bereitet man, abgesehen von der direkten Prüfungsvorbereitung, die Vorlesungen wöchentlich vor bzw. nach? | |
| Wie viele Stunden lernt man in der Vorbereitungszeit auf die Prüfungen durchschnittlich am Tag? | |

Vielen Dank für die Teilnahme an dieser Umfrage!

Vielen Dank für die Bereitschaft zur Teilnahme an dieser Befragung. Dieser Fragebogen dient ausschließlich einem wissenschaftlichen Zweck. Alle erhobenen Daten werden selbstverständlich anonym und vertraulich behandelt.

| | | |
|---|---|---|
| Welches Geschlecht haben Sie? | Männlich ☐ | Weiblich ☐ |
| Haben Sie Ihr Vordiplom bereits bestanden? | Ja ☐ | Nein ☐ |
| Wie viele Geschwister haben Sie? | | |
| In welchem Jahr wurde Ihre Mutter geboren? | | |
| In welchem Jahr wurde Ihr Vater geboren? | | |
| Wie viele Stunden Schlaf benötigen Sie, um sich fit zu fühlen? | | |
| Wie groß sind Sie (in cm)? | | |
| Wie hoch ist ungefähr Ihr monatliches Einkommen (in Euro) **vor** Abzug fixer Kosten (Lohn, Taschengeld, BAföG, …)? | | |
| Wie viel Geld (in Euro) steht Ihnen monatlich zur freien Verfügung? | | |
| Wie viel Miete bezahlen Sie monatlich (Euro)? | | |
| Wie weit entfernt von der Universität wohnen Sie (km)? | | |
| Wie viele Quadratmeter Wohnfläche stehen ausschließlich Ihnen zur Verfügung? | | |
| Wie viele Stunden jobben Sie regelmäßig während des Semesters pro Woche? | | |
| Wie viel Zeit (in Stunden) nehmen Sie sich wöchentlich für Hobbys (Sport, Musik, …)? | | |
| Wie viele Stunden arbeiten Sie pro Woche ehrenamtlich (Rotes Kreuz, Sportverein, …)? | | |
| An wie vielen Tagen (pro Semester) engagieren Sie sich außerhalb des Studiums im engeren Sinn für die Universität (StuRa, AStA, ...)? | | |
| Wie viele Stunden bereiten Sie, abgesehen von der direkten Prüfungsvorbereitung, die Vorlesungen wöchentlich vor bzw. nach? | | |
| Wie viele Stunden lernen Sie in der Vorbereitungszeit auf die Prüfungen durchschnittlich am Tag? | | |

Wenn Sie sich mit **anderen Studenten vergleichen**, wie sehr würden Sie dann folgenden Aussagen zustimmen? Beziehen Sie sich dabei bitte auf Ihre **derzeitige Situation**!

| | Stimme voll und ganz zu | Stimme zu | Stimme eher zu | Teils teils | Stimme eher nicht zu | Stimme nicht zu | Stimme über-haupt nicht zu |
|---|---|---|---|---|---|---|---|
| Ich verwende während des Semesters viel Zeit auf die Vor- und Nachbereitung von Vorlesungen | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| Mein tägliches Lernpensum ist in der Vorbereitungszeit auf die Prüfungen größer als bei anderen Studenten | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| Während des Semesters jobbe ich überdurchschnittlich viel | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| Für einen Studenten ist mein monatliches Einkommen (Lohn, Taschengeld, ...) relativ hoch (vor Abzug der Fixkosten) | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| Für einen Studenten habe ich sehr viel Geld zur freien Verfügung | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| Mein ehrenamtliches Engagement für die Universität (StuRa, AStA, ...) ist groß | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| Mein ehrenamtliches Engagement außerhalb der Universität (Rotes Kreuz, Sportverein, …) ist groß | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| Ich habe eine weite Anfahrt von meinem Wohnort zur Universität | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| Ich verwende überdurchschnittlich viel Zeit für Hobbys (Sport, Musik, …) | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| Ich bezahle eine vergleichsweise hohe Miete | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| Ich habe sehr viel Wohnraum für mich allein | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| Im Vergleich zu anderen habe ich viele Geschwister | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| Der Altersunterschied meiner Eltern ist groß | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| Ich benötige viel Schlaf, um mich fit zu fühlen | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |
| Ich bin überdurchschnittlich groß | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ | ☐ |

Beantworten Sie jetzt bitte die Fragen nochmals: Geben Sie jetzt **Ihre Einschätzung** ab, welche Werte für einen **durchschnittlichen Studenten** gelten:

| | |
|---|---|
| Wie hoch ist ungefähr das monatliche Einkommen in Euro **vor** Abzug fixer Kosten (Lohn, Taschengeld, BAföG, …)? | |
| Wie viel Geld (in Euro) steht monatlich zur freien Verfügung? | |
| Wie viel Miete wird monatlich bezahlt (Euro)? | |
| Wie weit ist die Entfernung von der Universität zum Wohnort (km)? | |
| Wie viele Quadratmeter Wohnfläche stehen ausschließlich zur eigenen Verfügung? | |
| Wie viele Stunden wird während des Semesters pro Woche ungefähr gejobbt? | |
| Wie viel Zeit (in Stunden) wird wöchentlich für Hobbys (Sport, Musik, …) verwendet? | |
| Wie viele Stunden pro Woche wird ehrenamtlich gearbeitet (Rotes Kreuz, Sportverein, …)? | |
| An wie vielen Tagen (pro Semester) erfolgt ein Engagement für die Universität außerhalb des Studiums im engeren Sinn (StuRa, AsTa, etc.)? | |
| Wie viele Stunden bereitet man, abgesehen von der direkten Prüfungsvorbereitung, die Vorlesungen wöchentlich vor bzw. nach? | |
| Wie viele Stunden lernt man in der Vorbereitungszeit auf die Prüfungen durchschnittlich am Tag? | |

Vielen Dank für die Teilnahme an dieser Umfrage!